First upper limits on the 21-cm signal power spectrum of neutral hydrogen at $z = 9.16$ from the LOFAR 3C 196 field

E. Ceccotti[1,2]⋆, A. R. Offringa[3,1], F. G. Mertens[4], L. V. E. Koopmans[1], S. Munshi[1], J. K. Chege[1,3],
A. Acharya[5], S. A. Brackenhoff[1], E. Chapman[6], B. Ciardi[5], R. Ghara[7], S. Ghosh[1], S. K. Giri[8,9],
C. Höfer[1], I. Hothi[10], G. Mellema[11], M. Mevius[3,1], V. N. Pandey[3] and S. Zaroubi[12,1,5]

[1] *Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands*
[2] *INAF – Istituto di Radioastronomia, Via P. Gobetti 101, 40129 Bologna, Italy*
[3] *ASTRON, PO Box 2, 7990 AA Dwingeloo, The Netherlands*
[4] *LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Université, F-75014 Paris, France*
[5] *Max-Planck Institute for Astrophysics, Karl-Schwarzschild-Straße 1, 85748 Garching, Germany*
[6] *School of Physics and Astronomy, The University of Nottingham, University Park, Nottingham, NG7 2RD, UK*
[7] *Department of Physical Sciences, Indian Institute of Science Education and Research Kolkata, Mohanpur, WB 741 246, India*
[8] *Van Swinderen Institute for Particle Physics and Gravity, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands*
[9] *Nordita, KTH Royal Institute of Technology and Stockholm University, Hannes Alfvéns väg 12, SE-106 91 Stockholm, Sweden*
[10] *Laboratoire de Physique de l'ENS, ENS, Université PSL, CNRS, Sorbonne Université, Universitée Paris Cité, 75005 Paris, France*
[11] *The Oskar Klein Centre, Department of Astronomy, Stockholm University, AlbaNova, SE-10691 Stockholm, Sweden*
[12] *Department of Natural Sciences, The Open University of Israel, 1 University Road, PO Box 808, Ra'anana 4353701, Israel*

ABSTRACT

The redshifted 21-cm signal of neutral hydrogen from the Epoch of Reionization (EoR) can potentially be detected using low-frequency radio instruments such as the Low-Frequency Array (LOFAR). So far, LOFAR upper limits on the 21-cm signal power spectrum have been published using a single target field: the North Celestial Pole (NCP). In this work, we analyse and provide upper limits for the 3C 196 field, observed by LOFAR, with a strong ≈80 Jy source in the centre. This field offers advantages such as higher sensitivity due to zenith-crossing observations and reduced geostationary radio-frequency interference, but also poses challenges due to the presence of the bright central source. After constructing a wide-field sky model, we process a single 6-hour night of 3C 196 observations using direction-independent and direction-dependent calibration, followed by a residual foreground subtraction with a machine learned Gaussian process regression (ML-GPR). A bias correction is necessary to account for signal suppression in the GPR step. Still, even after this correction, the upper limits are a factor of two lower than previous single-night NCP results, with a lowest 2σ upper limit of $(146.61 \, \mathrm{mK})^2$ at $z = 9.16$ and $k = 0.078 \, h \, \mathrm{cMpc}^{-1}$ (with $\mathrm{d}k/k \approx 0.3$). The results also reveal an excess power, different in behaviour from that observed in the NCP field, suggesting a potential residual foreground origin. In future work, the use of multiple nights of 3C 196 observations combined with improvements to sky modelling and ML-GPR to avoid the need for bias correction should provide tighter constraints per unit observing time than the NCP.

Key words: cosmology: dark ages, reionization, first stars – cosmology: observations – techniques: interferometric – methods: data analysis

1 INTRODUCTION

After cosmological recombination ($z \approx 1100$), hydrogen in the Universe became neutral and the Dark Ages began. During this era, matter density fluctuations grew under the influence of gravity, leading to the formation of the first stars and galaxies, marking the onset of the Cosmic Dawn (CD; $30 \gtrsim z \gtrsim 15$). Ultraviolet and X-ray radiation from these first objects initially heated and then ionized the surrounding intergalactic medium (IGM). This final major transition of the Universe, known as the Epoch of Reionization (EoR; $15 \gtrsim z \gtrsim 6$), marked the transformation of the IGM from largely neutral to largely ionized (e.g. Barkana & Loeb 2001; Loeb & Furlanetto 2013).

Our understanding of the EoR has advanced by indirect probes such as spectra of high-redshift quasars (e.g. Becker et al. 2001; Fan et al. 2006; Venemans et al. 2013; Eilers et al. 2018; Keller et al. 2024), Cosmic Microwave Background optical depth (e.g. Planck Collaboration et al. 2020), and Lyman-α emitters (e.g. Ouchi et al. 2010; Konno et al. 2014; Zheng et al. 2017; Taylor et al. 2021; Witstok et al. 2024). Recent breakthroughs by the James Webb Space Telescope have revealed an unexpectedly rich population of luminous objects at $z > 10$ (e.g. Bradley et al. 2023; Finkelstein et al. 2023; Carniani et al. 2024; McLeod et al. 2024; Chemerynska et al. 2024), challenging existing galaxy formation models (e.g. Arrabal

⋆ E-mail: emilio.ceccotti@inaf.it

Haro et al. 2023; Boylan-Kolchin 2023; Ferrara et al. 2023; Mason et al. 2023). However, uncertainties remain on the dominant ionizing sources, and these probes are unable to directly trace the reionization timeline. The most direct probe to map the evolution of the IGM during these early epochs is the redshifted 21-cm line from the hyperfine spin-flip transition of neutral hydrogen (see e.g. Furlanetto et al. 2006; Pritchard & Loeb 2012; Liu & Shaw 2020, for some reviews).

The 21-cm line emitted at $z > 6$ is observed at wavelengths longer than 1.5 m, and hence we need low-frequency radio instruments to detect it. Current instruments lack the sensitivity to directly image the 21-cm signal, aiming instead for a statistical detection of its spatial fluctuations by measuring its power spectrum. Instruments such as LOFAR[1] (van Haarlem et al. 2013; Mertens et al. 2020), MWA[2] (Lonsdale et al. 2009; Trott et al. 2020; Kolopanis et al. 2023), NenuFAR[3] (Zarka et al. 2012; Munshi et al. 2024), GMRT[4] (Paciga et al. 2013; Gupta et al. 2017), and HERA[5] (DeBoer et al. 2017; HERA Collaboration et al. 2023) have set increasingly stringent upper limits on the 21-cm power spectrum at different redshifts and scales, both at the CD and EoR. The upcoming SKA-Low[6] (Dewdney et al. 2009; Koopmans et al. 2015) is expected to have the sensitivity to make a direct detection and image the 21-cm signal (e.g. Giri et al. 2018a,b; Bianco et al. 2024).

All the aforementioned instruments face many challenges in detecting the 21-cm signal, which is a few orders of magnitude fainter than the astrophysical foregrounds, namely extra-galactic and Galactic emission. Because these foregrounds have smooth frequency spectra, being dominated by synchrotron radiation at low-frequencies, they can be separated from the 21-cm signal, which fluctuates rapidly, being a hyperfine spectral line (Jelić et al. 2008; Bowman et al. 2009; Bernardi et al. 2009, 2010). Foreground mitigation can be done with different techniques, such as foreground avoidance, where the Fourier modes dominated by foreground emission are directly discarded (e.g. Parsons & Backer 2009; Parsons et al. 2012; Thyagarajan et al. 2015), and foreground subtraction, where sky emission is subtracted from the data by accurate modelling (e.g. Patil et al. 2017; Mertens et al. 2020). However, radio interferometers are intrinsically chromatic and introduce spectral structures to the foregrounds, resulting in 'mode-mixing' that makes the signal separation harder (Morales et al. 2012, 2019). Errors can also be caused by inaccurate calibration (Barry et al. 2016; Beardsley et al. 2016; Patil et al. 2016; Ewall-Wice et al. 2017; Mazumder et al. 2022; Gorce et al. 2023; Ceccotti et al. 2025), polarization leakage (Jelić et al. 2010; Spinelli et al. 2018; Cunnington et al. 2021), ionospheric effects (Koopmans 2010; Vedantham & Koopmans 2016; Mevius et al. 2016; Brackenhoff et al. 2024), inaccurate primary beam model (Gehlot et al. 2021; Chokshi et al. 2024), gridding of visibilities (Offringa et al. 2019b), as well as mutual coupling (Kern et al. 2020; Kolopanis et al. 2023; Rath et al. 2024) and low-level radio-frequency interference (RFI; Offringa et al. 2019a; Wilensky et al. 2019).

By addressing many of these challenges, the LOFAR-EoR Key Science Project (KSP) has made continuous progress in recent years.

The KSP focuses on two main sky fields: the North Celestial Pole (NCP) field and the 3C 196 field, both of which are relatively cold spots in the Milky Way (Bernardi et al. 2010). The two fields have contrasting characteristics, some of which are described by Yatawatta et al. (2013). The main difference is the presence of a very bright FR-II source at the centre of the 3C 196 field, which can help the calibration by providing a high signal-to-noise ratio but may leave residuals after subtraction. Such a bright source is absent in the NCP field. Another difference is the elevation: with a declination of approximately $48°$, 3C 196 passes close to the zenith, whereas the NCP has a fixed elevation of $53°$. Consequently, the thermal noise in the 3C 196 direction is expected to be lower than in the NCP direction because the gain of the LOFAR dipole beam is maximum at zenith. On the other hand, the NCP field can be observed throughout the year and with a circular uv-coverage, as its pointing direction is fixed and parallel to the Earth rotation axis. This also introduces another difference, related to geostationary RFI (i.e. RFI sources that do not move with respect to the array), whose signals accumulate coherently in the NCP field, making it more challenging to remove (Munshi et al. 2025a). Being closer to the Galactic plane than the NCP, the 3C 196 field exhibits bright linearly polarized structures, which can again contaminate the 21-cm power spectrum if not properly accounted for (Jelić et al. 2015). This presents a challenge, as polarized emission is primarily attributed to diffuse Galactic emission and is affected by Faraday rotation, which are currently not accounted for in the LOFAR-EoR 21-cm signal processing pipelines. Nevertheless, Asad et al. (2015) demonstrated that such contamination remains at the level of the expected EoR signal, or below it if mildly subtracted, making it negligible for the current upper limits.

Both fields have been observed for thousands of hours, providing the sensitivity needed to detect the 21-cm signal in the absence of systematics. However, most of the calibration and analysis efforts have focused on the NCP field, where an extensive sky model was made by Yatawatta et al. (2013) at 150 MHz. This model was used by Patil et al. (2017) to get the first upper limits in the redshift range $z = 7.9–10.6$ from 13 h of data. The wide-field sky model was used for the initial calibration and then split into multiple directions to solve for direction-dependent (DD) station gains using the SAGE-CAL[7] code (Yatawatta 2016). The foreground subtraction was then performed by applying the DD solutions to the sky model, and the residual foreground emission was dealt with Generalized Morphological Component Analysis (GMCA; Bobin et al. 2007, 2008; Chapman et al. 2013). Using similar but improved processing pipeline on 141 h of data (12 nights) and replacing GMCA with the Gaussian process regression (GPR; Mertens et al. 2018), Mertens et al. (2020) have been able to set deeper upper limits at $z \approx 9.1$ on the spherical power spectrum, namely $\Delta_{21}^2 < (72.86\,\mathrm{mK})^2$ at $k = 0.075\,h\,\mathrm{cMpc}^{-1}$ (with $\mathrm{d}k/k \approx 0.3$).

These results already set some constraints on the physics of the EoR (e.g. Ghara et al. 2020; Mondal et al. 2020; Greig et al. 2021), but the upper limits were systematics limited rather than thermal noise limited. Consequently, increasing the observing time does not lower the upper limits if the excess power is not incoherent in time like the thermal noise. Understanding the origin of this excess is therefore crucial for mitigating or avoiding it. Recent studies on the NCP field have shown that ionospheric effects are unlikely to be the primary contributors (Gan et al. 2022; Brackenhoff et al. 2024), although bright, distant sources such as Cassiopeia A (Cas A) and Cygnus A (Cyg A) might play a role, especially in combination with incorrect

---

[1] Low-Frequency Array, https://www.astron.nl/telescopes/lofar

[2] Murchison Widefield Array, http://www.mwatelescope.org

[3] New Extension in Nançay Upgrading LOFAR, https://nenufar.obs-nancay.fr

[4] Giant Metrewave Radio Telescope, https://www.gmrt.ncra.tifr.res.in

[5] Hydrogen Epoch of Reionization Array, https://reionization.org/

[6] Square Kilometre Array, https://www.skao.int/en/explore/telescopes

[7] https://sagecal.sourceforge.net/

primary beam models and DD gain errors (Gan et al. 2022; Ceccotti et al. 2025; Brackenhoff et al. 2025). Despite these effects, GPR (Mertens et al. 2018) has demonstrated competitive performance even in the presence of such an excess (Hothi et al. 2021), although this methodology might suffer from biases and signal suppression, as highlighted by Kern & Liu (2021). To address these effects, a machine learning-enhanced GPR framework was developed by Mertens et al. (2024), incorporating parametrized covariance functions derived from simulations. This framework has been successfully tested on both simulations (Acharya et al. 2023) and real LOFAR data (Acharya et al. 2024). The better understanding of the excess power has led to new upper limits at $z \approx$ 8.3, 9.1, and 10.1 from the NCP field (Mertens et al. 2025), deeper by a factor of 2−4, depending on the k-mode, at $z \approx$ 9.1 compared to Mertens et al. (2020), using nearly the dame data sets.

While most efforts have focused on the NCP field, investigating the second deep field, centred on 3C 196, could provide new insights into the origin of the excess power. Low-band observations of 3C 196 in the 56−70 MHz range (i.e. $19 \lesssim z \lesssim 24$) were analysed by Gehlot et al. (2018), revealing a significant excess power across all baseline lengths. Similar to the NCP case, potential sources for this excess have been proposed, such as an incomplete sky model, ionospheric effects, and calibration errors, but no extensive investigation has been carried out in the EoR frequency window. Additionally, it has been demonstrated that comparing results from different fields is crucial for better understanding the sources of systematics (e.g. Trott et al. 2020; Rahimi et al. 2021; Abdurashidova et al. 2022).

In this paper, we present the first 21-cm signal power spectrum upper limits at $z \approx$ 9.1 from LOFAR observations of the 3C 196 field. As discussed earlier, the 3C 196 field poses different challenges compared to the NCP, particularly due to the presence of a bright central source, which makes its processing and analysis more complex. Building on the experience gained by the LOFAR-EoR team from NCP observations, we developed a processing pipeline, tailored for this field, incorporating the latest software and techniques. With this pipeline, we analysed a single 6-hour night of observed data resulting in a 21-cm signal power spectrum independent from the NCP field.

The paper is organized as follows. In Section 2, we introduce the 3C 196 dataset, which is used to extract a wide-field sky model and estimate the 21-cm upper limits. In Section 3, we describe the process of obtaining the sky model for the field, with several processing steps mirroring those used in the new EoR pipeline, detailed in Section 4. In Section 5, we discuss the GPR method and present the resulting power spectra. In Section 6, we focus on the validation of the 21-cm signal processing pipeline, particularly the GPR technique. The results and final upper limits derived from the 3C 196 field are presented in Section 7. Finally, in Section 8, we summarise the findings and draw the conclusions. Throughout this paper, we used a flat ΛCDM cosmology consistent with the Planck Collaboration et al. (2016) results, similar to Mertens et al. (2020).

2 LOFAR-HBA OBSERVATIONS OF THE 3C 196 FIELD

LOFAR (van Haarlem et al. 2013) is a low-frequency radio interferometer constituted by stations spread across the Netherlands and Europe. It currently consists of 24 core stations (CS; max baseline $\approx$ 4 km) and 14 remote stations (RS; max baseline $\approx$ 120 km) in the Netherlands, forming the Dutch array, and 14 international stations (max baseline $\approx$ 2000 km) spread across Europe. Every station is organized into two types of antenna sets observing different frequency ranges: Low-Band Antennas (LBA) for 10−90 MHz and High-Band Antennas (HBA) for 110−250 MHz. Each CS has two HBA substations (HBA0 and HBA1) that can operate independently ('HBA Dual' mode) for a better uv-coverage. Each station acts as a phased array that can track a phase centre on the sky via beam-forming inside the dipole beam.

Since its first observations in 2011, the LOFAR-EoR KSP has focused on observing two main windows on the sky using the HBA system: the North Celestial Pole (Patil et al. 2017; Mertens et al. 2020, 2025) and the 3C 196 fields Gehlot et al. (2018). Compared to the NCP (see Yatawatta et al. 2013), the 3C 196 field presents advantages and disadvantages. Firstly, with a flux density of about 83 Jy at 150 MHz (Scaife & Heald 2012), 3C 196 requires a very accurate model to minimize its residuals during the foreground subtraction step (see Section 4.3). Once such a model is in place, having a bright source in the centre makes the direction-independent calibration easier because 3C 196 dominates the observed visibilities. Another challenge is that, while the NCP phase centre is fixed in the sky and is observable by LOFAR at the same elevation throughout the year, 3C 196 moves due to the Earth rotation and is observable only for half of the year. A benefit is that it can be observed close to the zenith, having a declination of 48.2° (whereas LOFAR core is at a latitude of $\approx$53°), where the dipole beam suppression is more limited compared to the NCP. This results in a higher sensitivity (approximately a factor of 1.5 more than for the NCP field). The fixed pointing direction of the NCP also leads to geostationary RFI to add up coherently, having the same fringe speed as sources in the target field, while this is not the case for a tracking field such as the one centred on 3C 196 (Munshi et al. 2025a).

2.1 Selected data set

In this work, we analysed a single night of 6 h LOFAR-HBA observations taken in 2014 (with elevation ranging from 61° to 85°, peaking at the midpoint of the observation). The recorded data consists of 380 sub-bands (SBs) spanning 115–189 MHz, where each SB is 195.3 kHz wide, with a frequency resolution of 3.05 kHz (i.e. 64 channels per SB) and an integration time of 2 s. However, because of the poly-phase filter (Brentjens & Mol 2018), the four edge channels of each SB are affected by aliasing and were flagged before starting any processing. This reduces the SB width to 183.1 kHz. The data were flagged with AOFLAGGER (Offringa et al. 2012), and averaged to 5 channels (36.6 kHz each) per SB and 4 s for archival purposes. The data were stored uncompressed because DYSCO compression (Offringa 2016; Chege et al. 2024) was not yet implemented at that time. The resulting data set is the starting point of this work and will be referred to as the initial data set. Further observational details are reported in Table 1. All the processing described in this paper has been performed on the 'Dawn' high-performance GPU computing cluster (Pandey et al. 2020) at the University of Groningen.
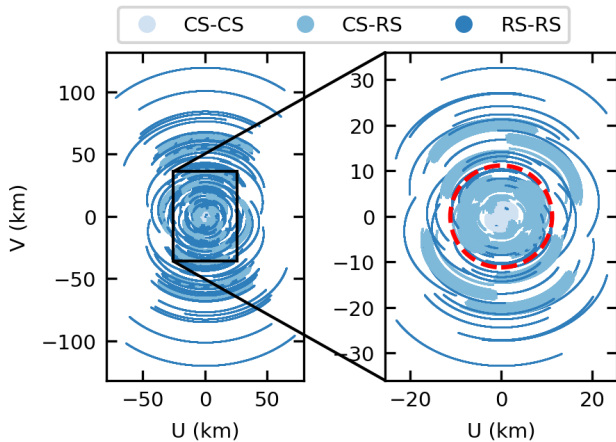
3 SKY MODELLING

To calibrate our data, a sky model of the 3C 196 field is required. While for a direction-independent (DI) calibration just using the bright 3C 196 source might be enough, for the direction-dependent calibration and subtraction required before the power spectrum estimation, we need an extensive, wide-field model of the surrounding field. In the following, we describe the processing steps used to produce such a sky model.

We used the same data set presented in Section 2.1, but we selected 246 SBs spanning the 120–168 MHz frequency range. The choice of

Table 1. Observational details of the initial 3C 196 data set used in this work.

Parameter	Value
Telescope	LOFAR HBA
Project code	LC3_028
Observation ID	L253456
Antenna configuration	HBA Dual Inner
Number of stations	61 (CS + RS)[a]
Phase centre (J2000):	
Right Ascension	$08^h13^m36^s07$
Declination	$+48°13'02''58$
Obs. start time (UTC)	2014 Dec 02; 23:58:39
Frequency range:	
Full data set	115.0 − 189.1 MHz
Redshift bin	134.2 − 147.1 MHz
Duration of observation	6 h
Time resolution	4.0 s
Frequency resolution	36.6 kHz

[a] RS310 did not participate in this observation.

Figure 1. Baseline coverage in the *uv*-plane for the LOFAR-HBA observation of the 3C 196 field, before (left, zoomed out) and after (right, zoomed in) the flagging performed in the pre-processing step (see Section 3.1 and Section 4.1). The antenna pairs are shown in different colours: core-to-core station baselines in light blue, core-to-remote (and remote-to-core) station baselines in medium blue, and remote-to-remote baselines in dark blue. In the right panel, the red circle marks the 5000λ limit that is usually applied during the DI calibration of the NCP field.

not using the full 75 MHz bandwidth comes from two main reasons: (i) the band edges are mostly affected by RFI, and (ii) to speed up calibration and imaging. Because the data set has already been flagged before archiving, we can start the processing directly with the DI calibration, described in the next section.

3.1 Baseline selection

In general, when we are interested in a single, dominant source at the phase centre, DI calibration can be performed with all the available baselines, resulting in the maximum signal-to-noise ratio for the station gains. In this case, most of the DD effects can be neglected. In our case, we want to extract a model of a few degrees in spatial extent from the target field, which means that we have to minimize the DD effects, such as ionospheric distortion and smearing of distant

sources (e.g. Patil et al. 2016; Vedantham & Koopmans 2015). This is done by removing the longest baselines, applying an outer *uv*-cut. By removing only the longest baselines, most stations still have sufficient baselines to obtain accurate gain solutions. The challenge here is that we still want long baselines to get a sky model with a high enough spatial resolution. We found that using baselines shorter than ≈30 km (i.e. ≈15 000λ at 150 MHz) is a good compromise between minimizing ionospheric distortion, given a diffractive scale of about 15 km for our observed night, and maximising spatial resolution, whose full width at half maximum (FWHM) is approximately 14 arcsec. Instead of cutting in the *uv*-plane, we removed all the baselines that had a physical length larger than 30 km. This avoids having gaps in the *uv*-tracks for some baselines during the time synthesis, due to the ellipticity of *uv*-tracks. We also removed a few RS that resulted in just one baseline, which is not enough to obtain reliable gain solutions. Fig. 1 shows the *uv*-coverage before (left panel, zoomed out) and after such baseline selection (right panel, zoomed in). Furthermore, we also removed the intra-station baselines (HBA0-HBA1) of the core array, because of possible correlated RFI generated inside the shared electronics cabinet. This baseline selection is applied to the data set before DI calibration for the sky modelling part and is part of the pre-processing step of the EoR pipeline, later described in Section 4.1. Additionally, a lower *uv*-cut of 50λ is applied throughout the entire processing to avoid poor *uv*-coverage and strong unmodelled diffuse emission (Patil et al. 2017).
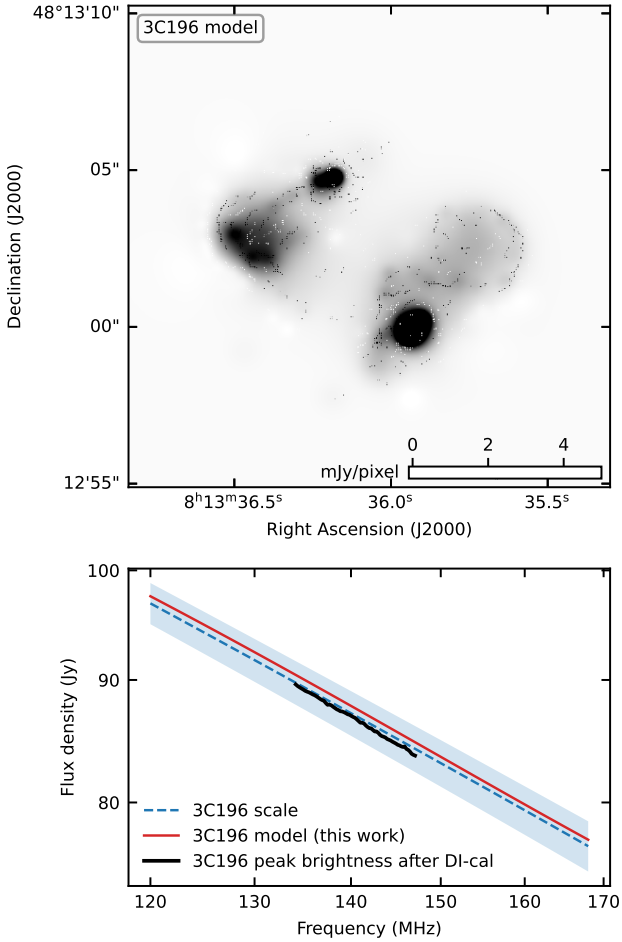
3.2 Direction-independent calibration

The DI calibration is performed by using a high-resolution, multiscale LOFAR-HBA model of 3C 196, shown in the top panel of Fig. 2. 3C 196 is a compact source approximately 10 arcsec in size, so LOFAR international stations were used to make such a model, which is described by 1812 components (634 point sources and 1178 Gaussians). The total flux agrees with the Scaife & Heald (2012) scale, as shown in the bottom panel of Fig. 2. To process and calibrate the data, we used DP3[8] (Default Pre-Processing Pipeline; van Diepen et al. 2018), which collects many tools to perform different operations. DI calibration was divided in two steps, similar to the 21-cm signal processing pipeline that we will describe in Section 4: (i) a spectrally smooth calibration, and (ii) a bandpass calibration.

The main goal of the spectrally smooth calibration is to correct for high-temporal effects in the station gains. Using a spectral constraint helps keep the solutions stable over frequency while solving at the highest time resolution possible. This initial calibration was performed with the DDECAL tool (for an extensive description, see Brackenhoff et al. 2025), which applies a Gaussian smoothing kernel at every iteration to enforce spectral smoothness in the solutions. In our case, a kernel of 1 MHz width was used. We solved both amplitudes and phases of the full Jones matrix (a 2 × 2 complex matrix describing the effect of a station with orthogonal polarizations) for each SB (183.1 kHz) and each time interval (4 s). It is important to solve for a full Jones matrix (i.e. four complex gains) to correct any polarization leakage. Sometimes leakage from one Stokes mode to another can occur because of the primary beam, as we will show in Section 4.3. During calibration, the primary beam model response is applied to the 3C 196 model to obtain apparent visibilities. By using the beam model throughout the processing, we can later extract intrinsic flux values of the entire field. In DP3, the primary beam model

---

[8] https://dp3.readthedocs.io/

Figure 2. Rendered model of the 3C 196 high-resolution model at 150 MHz (top) and the flux density over the 120–168 MHz range (bottom). The flux estimated by Scaife & Heald (2012) is plotted with the dashed blue line, with the 1σ uncertainties as the blue shaded area. The flux of the 3C 196 model used in this work is shown with the red line. We also show the peak brightness of the source at each frequency of the z-bin, extracted from the dirty images after the DI-calibration of the EoR pipeline (Section 4.2), with the black line.

includes both the element beam, which describes the directional sensitivity by combining the dipole beams in a single tile, and the array factor, which characterises the actual field of view of a station (van Haarlem et al. 2013). These two beams are independently calculated and can be multiplied to obtain the full LOFAR primary beam. The array factor is negligible in our case, as the 3C 196 is compact and in the centre of the field, making the array factor one at its position. However, since the field is not observed at zenith, the element beam is time-dependent and not unity. The response is approximately 90 per cent at the meridian. After calibration, the gain solutions were applied to the data.

After correcting for fast effects, the fine frequency response of the stations must be calibrated. For this bandpass calibration, we used the GAINCAL algorithm of DP3 for finding solutions (Mitchell et al. 2008; Salvini & Wijnholds 2014). We solved for each channel (36.6 kHz) over a time interval of 1 h to average out the temporal behaviour of the instrument and obtain reliable spectral corrections. To reduce the number of degrees of freedom, we solved only for the diagonal elements of the Jones matrix. The bandpass solutions

were used to subtract 3C 196 from the data set, before applying them to the residual visibilities. Removing the central source is essential for the next imaging step, where the wide-field sky model will be extracted. Deconvolution struggles to handle the sidelobes of such a bright source, but our 3C 196 model, which was built using very long baselines, is more spatially accurate than what imaging can reproduce from our data set, which includes Dutch baselines only. This allowed us to better remove the 3C 196 sidelobes and achieve a more complete and accurate sky model for the rest of the field. The calibrated data were finally corrected for the element beam in the visibility space by scaling the data to have the correct flux density at the phase centre. To extract an intrinsic model of the 3C 196 field, we applied the more rapidly spatially varying beam (i.e. the array factor term) during the imaging step, which is described in the next section.
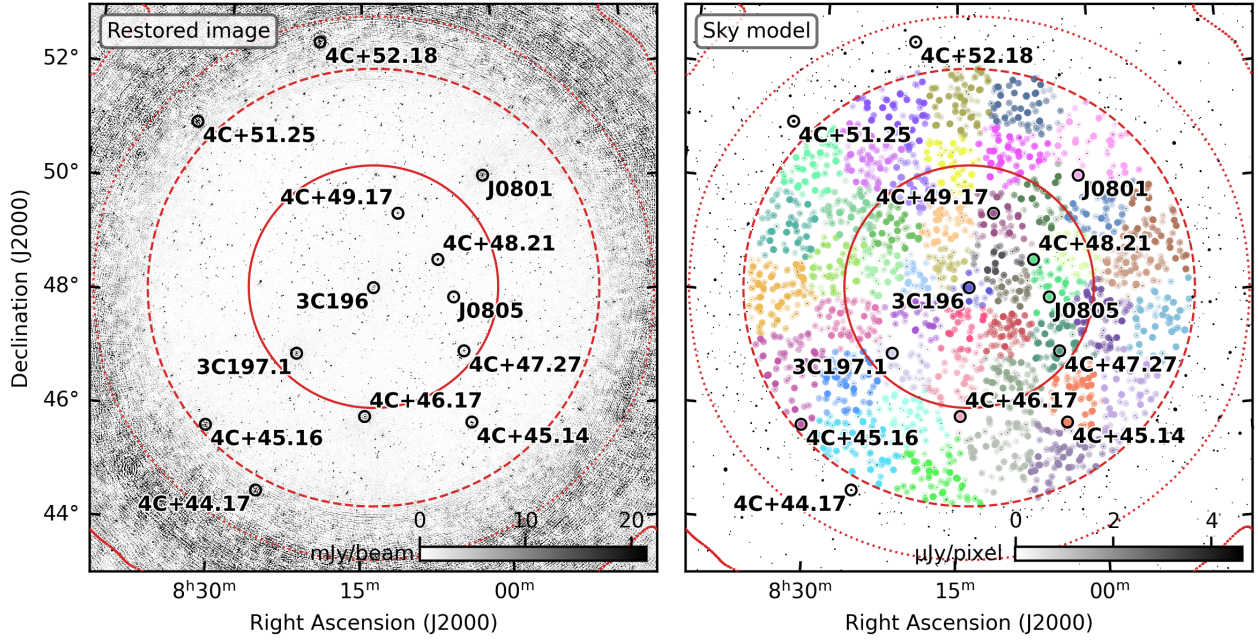
3.3 Imaging

The sky model of the 3C 196 field was made by taking the multi-scale clean components (Gaussian and point sources) from the calibrated data using the WSCLEAN imager software (Offringa et al. 2014). We used its multi-frequency, joined-channel deconvolution (Offringa & Smirnov 2017) to accurately model the spectral features of the sources. This method splits the full bandwidth into multiple channels, while peak finding of the clean components is performed on a frequency-integrated image, which has a higher signal-to-noise ratio. Cleaning is then performed in each output channel at the identified component positions, and spectral smoothness can be enforced by fitting an ordinary or logarithmic polynomial function. In our case, we used 12 output channels, each 4 MHz wide, and fitted an ordinary, third-order polynomial function with four terms. Although sources at low frequencies are primarily dominated by synchrotron emission and thus have a power-law spectral energy distribution, using a logarithmic function for fitting often results in incorrect spectral index values due to limited bandwidth or significant systematics (e.g. Offringa et al. 2016). While ordinary polynomial functions are generally more stable, they are challenging to use at different frequencies. However, since the sky model of the 3C 196 field will be used in the LOFAR-EoR pipeline within the frequency range from which it was extracted, this is not an issue.

To ensure the best quality for each output channel, we did not apply multi-frequency weighting[9]. We used the W-gridder algorithm (Arras et al. 2021) for gridding the visibilities. Each output channel was then weighted with a Briggs weighting scheme with a robust parameter of −1, resulting in a synthesised beam with a full width at half maximum (FWHM) of 13 × 18 arcsec at the lowest frequency channel (centred on 122 MHz). To achieve a more circular beam, we applied a Gaussian taper of 20 arcsec in size to the gridded uv-plane. Using a pixel size of 3 arcsec was sufficient to sample the resulting synthesised beam.

Alongside the multi-frequency deconvolution, we also used the multi-scale CLEAN algorithm (Offringa & Smirnov 2017) to accurately model the spatial structures of the sources. The multi-scale deconvolution is useful for cleaning resolved sources, representing source models as a summation of basis functions (Gaussian or tapered quadratic) of different sizes and point-like components. Since we are not focusing on a single source and our resolution is degraded by the Gaussian taper, we limited the multi-scale algorithm to four spatial scales, namely 0 (delta function), 1.4, 2.7, and 5.4 arcmin.

---

[9] More details in https://wsclean.readthedocs.io/en/latest/mf_weighting.html.

Figure 3. The left panel shows the frequency-integrated (120–168 MHz) restored image of the 3C 196 field, where 3C 196 has been subtracted and the full primary beam correction applied. The image noise is $\sigma = 0.8$ mJy/beam. The model image resulting from the imaging process is shown in the right panel, where the sources of the 47 clusters are highlighted by different colours. Both images have a pixel resolution of 3 arcsec and a field of view of $10° \times 10°$. The 1, 7.5, and 50 per cent levels of the primary beam intensity are shown with the red contours. The 7.5 per cent level roughly corresponds to the 3.9° radius that we used to select the sky model components. The brightest 3C and 4C sources are also indicated by black circles, in addition to J080135.35+500943.9 (shortened to J0801) and J080508+480151 (shortened to J0805) that have approximately 11 and 2 Jy of total flux, respectively.

The cleaning was performed on the flat-noise images (i.e. before the spatially varying primary beam correction) to an initial threshold of 5σ, at which a pixel mask was made by using the auto-masking feature of WSCLEAN, finally cleaning to 1σ within that mask, where $\sigma = 0.15$ mJy beam^{-1} for the frequency-integrated image.

The spatially varying beam is then corrected from the deconvolved images and the clean model to obtain the intrinsic fluxes. The primary beam size and shape vary with frequency, having a narrower main lobe at higher frequencies. Sources that are near or in the predicted beam null can cause stability issues. For these reasons, we focused on imaging mainly the main lobe, which has a FWHM of approximately 4° at 150 MHz, resulting in an image size of 12k×12k pixels to cover a sky area of $10° \times 10°$.

The frequency-integrated continuum image of the 3C 196 field, with 3C 196 subtracted and after the full primary beam correction, is shown in the left panel of Fig. 3. The red contours indicate the primary beam intensity as a percentage of the peak (which is one at the phase centre) for the frequency-integrated data. This means that, for instance, the 1 per cent level corresponds to a larger area at low frequencies and a smaller area at higher frequencies. In the image, outside the 7.5 per cent contour, the noise boost due to the beam correction becomes appreciable. The cleaning process also found components in this area, as shown in the right panel of Fig. 3. Cutting the image field of view at 10° was insufficient to exclude such sources, which can cause stability issues due to the low value of the beam response. In the next section, we describe the selection performed on such a raw clean model to obtain the final, cleaned sky model that is used in the 21-cm signal processing pipeline.
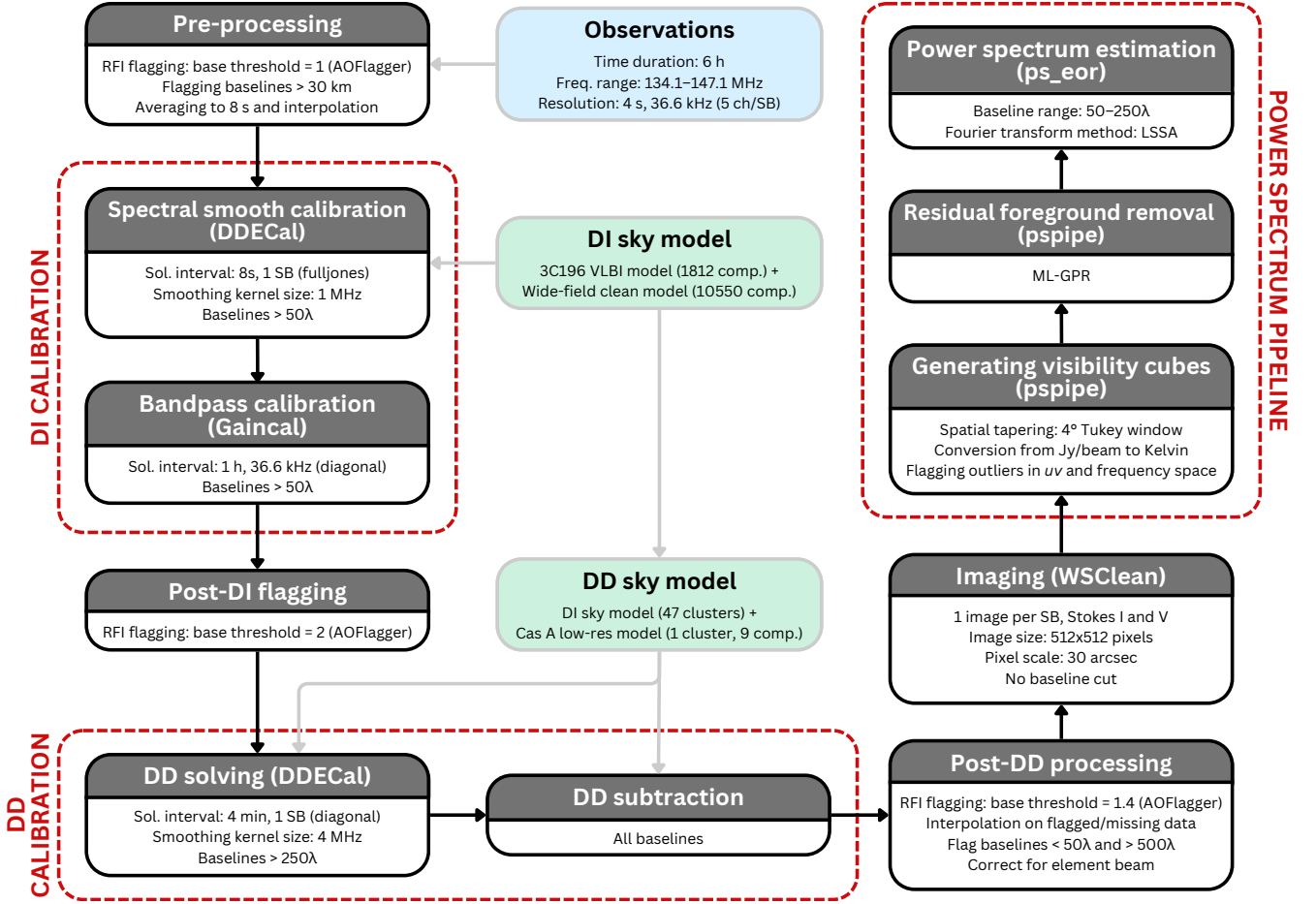
3.4 Sky model extraction and clustering

Selection and cleaning of the sky model sources were performed in three steps:

(i) We removed deconvolved residuals, which are visible in the position of 3C 196, visible in the restored image on the left panel of Fig. 3. This was done by cutting out all the cleaned components within a central aperture with a radius of 2 arcmin.

(ii) We only kept components within an aperture centred on 3C 196 with a radius of 3.9°, corresponding to the 7.5 per cent level of the frequency-integrated primary beam main lobe, shown in Fig. 3. The components selected in this way were not affected by the weaker response of the primary beam.

(iii) We removed deconvolution artefacts that resulted in isolated negative components, although the total intensity of these components was only −231 mJy at 150 MHz.

These operations resulted in a final sky model with 10 550 components, to which we added the 3C 196 model used in the DI calibration, bringing the total to 12 362 components. The resulting model is referred to as the 'DI sky model' because it is used for the DI calibration steps of the 21-cm signal processing pipeline.

This model is then split into multiple directions (or clusters) in the sky for direction-dependent (DD) calibration. Clustering was done using a modified K-means algorithm that considers angular distances instead of Euclidean ones. Therefore, due to projection effects, outer clusters are more extended than central ones. The K-means algorithm does not account for the flux density of clustered sources, so we validated that each direction had sufficient signal-to-noise ratio during DD solving. We aimed for a flux density greater than $5\sigma_{DD}$, where σ_{DD} is the noise level within a DD calibration

Figure 4. The 21-cm signal processing pipeline for the LOFAR 3C 196 field, from the pre-processing and DI calibration to the final power spectrum estimation.

gain interval, calculated using the radiometer equation:

$$\sigma = \frac{\text{SEFD}}{\sqrt{2\Delta t \Delta \nu}}, \qquad (1)$$

where SEFD is the system equivalent flux density, Δt is the time interval, and $\Delta \nu$ is the frequency interval. For DD gain solutions (described in Section 4.3), we used $\Delta t = 4$ min and a spectral smoothing kernel of 4 MHz, approximating $\Delta \nu$ with this value. With an average SEFD of 2835 Jy per visibility for our data set, we get $\sigma_{\text{DD}} \approx 65$ mJy, meaning each cluster needs at least 325 mJy.

We reached this flux level with our sky model by using 60 directions, but this resulted in small clusters containing only a few tens of components. Clusters should generally be $1-2°$ in size to ensure enough signal-to-noise for the solutions, while also reducing the spatially varying DD effects. In addition, having too many clusters also increases the degrees of freedom in the calibration, which raises both computational cost and the risk of absorbing unmodelled signal into the solutions. To address both the signal-to-noise and degree-of-freedom aspects, we aimed to ensure that each cluster contained more than 100 components and was approximately $1°$ in size, opting for 47 directions. This resulted in all clusters having more than 1 Jy total flux and more than 110 components (see Appendix A). Components within the same cluster are highlighted in the same colour in the right panel of Fig. 3. To this model, we finally add a low-resolution model

of Cas A[10] with nine components in a separate cluster. The resulting model is referred to as the 'DD sky model' from here onward.

4 3C 196 PROCESSING PIPELINE

The data processing for the 3C 196 field is based on an updated pipeline from the NCP processing (Mertens et al. 2020), where we replaced SAGECAL by DP3. This choice was motivated by the need for a well-integrated environment that provides tools for processing, flagging, and both DI and DD calibration, while also supporting sky models in the format output by WSCLEAN. These tools are the standard for processing LOFAR observations and are implemented in pipelines such as LINC[11] (LOFAR initial calibration; de Gasperin et al. 2019) and RAPTHOR[12]. A fully DP3-based 21-cm signal processing pipeline has been developed and used by Munshi et al. (2024) to set the first upper limits on the Cosmic Dawn signal from NenuFAR.

The 3C 196 processing pipeline is designed to account for both DI and DD effects while preserving the cosmological 21-cm signal. It consists of several stages: (1) pre-processing and RFI flagging,

---

[10] https://github.com/lofar-astron/prefactor/blob/master/skymodels/A-Team_lowres.skymodel
[11] https://linc.readthedocs.io/en/latest/index.html
[12] https://rapthor.readthedocs.io

(2) DI calibration, (3) DD solving and sky model subtraction, (4) imaging, (5) modelling and removal of residual foregrounds, and (6) power spectrum estimation. The pipeline is applied to the frequency range 134.2–147.1 MHz, corresponding to the redshift bin centred at $z = 9.16$. Steps (1) and (2) are similar to the processing for generating the sky model (Section 3). In the following, we provide a detailed description of each step of the 3C 196 21-cm signal pipeline, an overview of which is presented in Fig. 4. Step (5) involves applying the GPR method (Mertens et al. 2018) and is described in Section 5.

4.1 Pre-processing

The initial data set had already undergone partial pre-processing when the observations were taken in 2014, including averaging to 5 channels per SB and 4 s time intervals. However, we found residual RFI that had not been properly removed. To address this, we used AOFLAGGER with an updated RFI flagging strategy optimised for LOFAR-HBA data. AOFLAGGER uses an adaptive thresholding algorithm to automatically detect and flag corrupted data. During the initial pre-processing, we set the base threshold to the default value of 1, which corresponds to a flagger false-positives rate of 0.5 per cent (Offringa et al. 2013)[13]. This resulted in about 3.4 per cent of visibilities flagged. Additionally, we removed baselines longer than 30 km using the same method as described in Section 3.1 to minimize the effects of ionospheric distortions.

Following the RFI and baseline flagging, the data were then averaged to 8 s time intervals. This averaging step reduces the data volume, thereby decreasing computational costs in subsequent processing steps, without decorrelating the signal from the longer baselines. To further mitigate the impact of flagged or missing data, a Gaussian-weighted interpolation scheme was applied. This interpolation ensures that the introduced values are smooth and consistent with the surrounding data, preventing the generation of artificial spectral fluctuations (see Offringa et al. 2019a for details).

4.2 Direction-independent calibration

The DI calibration is performed similarly to Section 3.2, solving first for spectrally smooth gains on short time intervals and then for the high spectral resolution bandpass response on long time intervals. We used the DI sky model for both steps. Because the model is intrinsic by construction, we had to apply the full LOFAR primary beam model. For the spectrally smooth calibration, we solved for the high-temporal effects, using the DDECAL tool to apply a Gaussian smoothing kernel of 1 MHz width on the per-station solutions at every iteration. We solved for full Jones matrices with a time interval of 8 s (i.e. one integration time) and a frequency interval of 183.1 kHz (i.e. one SB). The resulting solutions are shown in Fig. B1. Both amplitudes and phases were applied to the data.

After the spectral smooth calibration, a bandpass calibration was performed using the GAINCAL tool to correct for the frequency response of the instrument. We solved for diagonal gains to reduce the number of free parameters, using a time interval of 1 h and a frequency interval of 36.6 kHz (i.e. one channel). The gain solutions are shown in Fig. B2. As with the spectral smooth calibration, we applied both amplitudes and phases to the data.

---

[13] For more details about the algorithm parameters, we refer to Offringa et al. (2010) and https://aoflagger.readthedocs.io

A second round of RFI flagging was performed on the DI calibrated data with a decreased sensitivity (an AOFLAGGER base threshold of 2) compared to the pre-calibration flagging. This step removed residual RFI that may have been left unflagged during the initial pre-processing, without strongly affecting the data by using a higher threshold. Only 0.9 per cent of the remaining data were flagged.

The dirty images of the DI calibrated data, made with WSCLEAN, are shown in the top panels of Fig. 5, with the all-sky image shown on the left, while a zoom-in of the central 3C 196 field is shown on the right, with the same size of Fig. 3. Both the continuum images are frequency-integrated within the redshift bin bandwidth (134.2–147.1 MHz). We subtracted 3C 196 to make the rest of the field visible. A Briggs weighting scheme with robust parameter of −2 and a uv-coverage Gaussian taper of 10 arcmin have been used for the all-sky image, while natural weighting and uv-range of 50–500λ have been set for the zoom-in. The all-sky image is dominated by sidelobes from the central field sources, which are still visible despite the use of the Gaussian taper, making even bright A-team sources hard to see. No point spread function (PSF) sidelobes of A-team sources are evident in the zoomed-in image, which is instead dominated by 3C and 4C sources. In fact, the 3C 196 field presents many more of these sources than the NCP, with flux densities higher than 1 Jy, both inside and outside the main primary beam lobe. We also calculated the standard deviation of the Stokes I gridded data for each uv-cell. This is shown in the left panel of Fig. 6 in SEFD units, between 50 and 500λ. The power seems uniformly spread over the uv-grid, with no clear directionality visible, which suggests that it comes from many sources (Munshi et al. 2025b), unlike the NCP field where Cas A and Cyg A dominate the sidelobe leakage into the central field (Mertens et al. 2018; Gan et al. 2022).
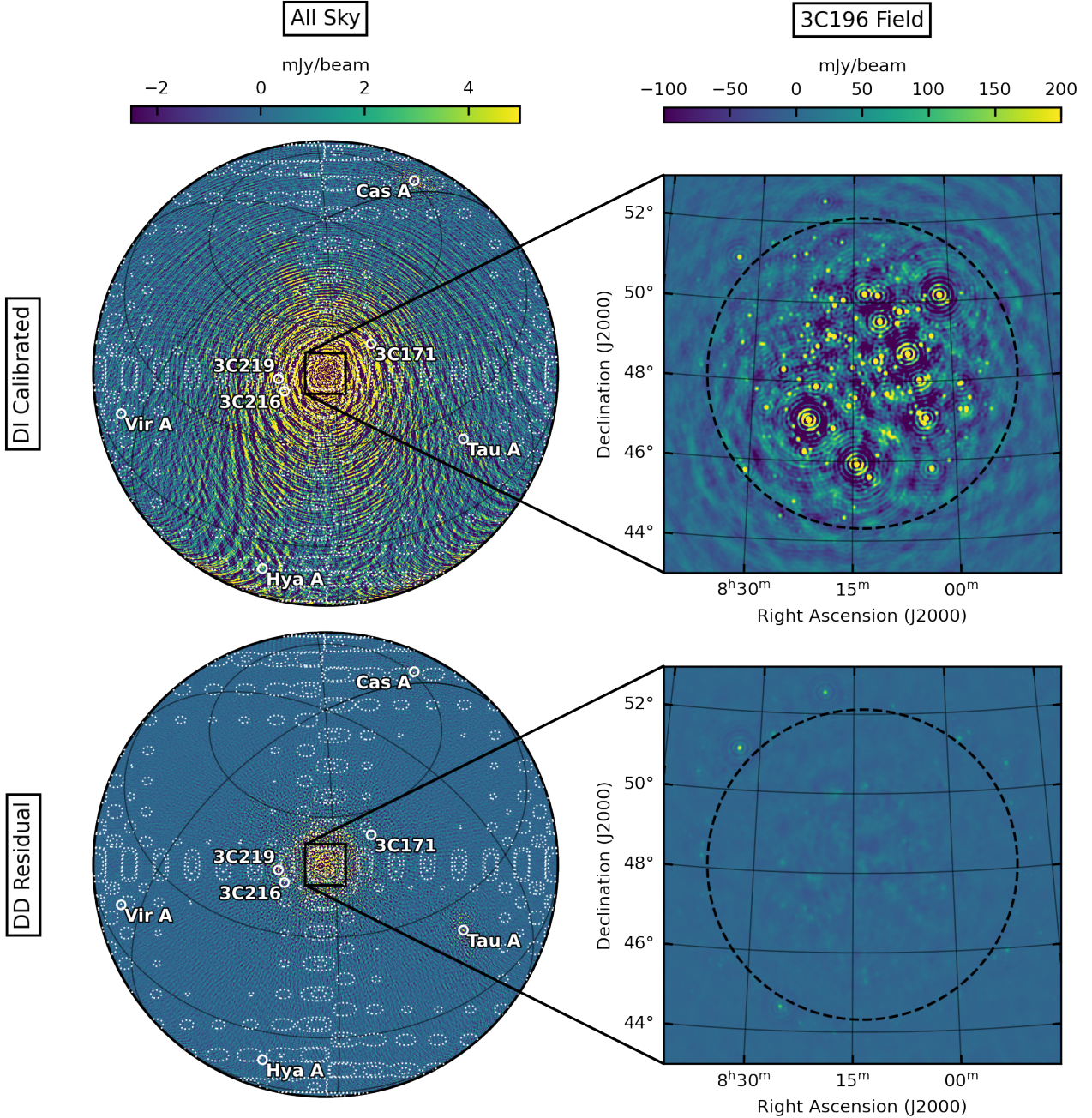
4.3 Direction-dependent calibration

The large field of view of LOFAR introduces significant DD effects due to the time-varying ionosphere and the imperfect knowledge of the primary beam. These effects are addressed by performing a DD calibration, where the sky model is divided into multiple directions (or clusters), and solutions are derived for each direction. Because of the bright sources present in the 3C 196 field, clusters of 1° were large enough to fulfil these conditions, resulting in a DD sky model with 47 directions as described in Section 3.4. In addition, we included a model of Cas A, which is the brightest A-team source in apparent flux, as shown with dashed lines in Fig. 7, but only during the first quarter of the observation. The source is at the edge of a primary beam sidelobe, plotted with dotted white lines in the left column of Fig. 5, and its elevation decreases over the night. The second brightest source is Taurus A (Tau A), which is also the closest to the main field. Because we do not have a well-tested model of it, we do not include Tau A in this work. In general, subtracting an incorrect source model can introduce more residual contamination than omitting it entirely. Instead, we let GPR handle the source sidelobes, which may lead to better results, as shown in the latest NCP analysis, where removing poorly modelled sky clusters improved the results (Mertens et al. 2025).

The DD calibration consists of two steps: (i) solving for the frequency-dependent gain solutions per source cluster, and (ii) subtraction of the sky model 'corrupted' by these gains. Thus, the gain solutions are not applied to the data but to the model. To find the DD solutions, we used the directional solving algorithm of DDECAL (Smirnov & Tasse 2015), which allowed us to apply a frequency smoothing kernel to the gains, similarly to the spectral smooth DI calibration. In this case, we used a Gaussian kernel with a 4 MHz
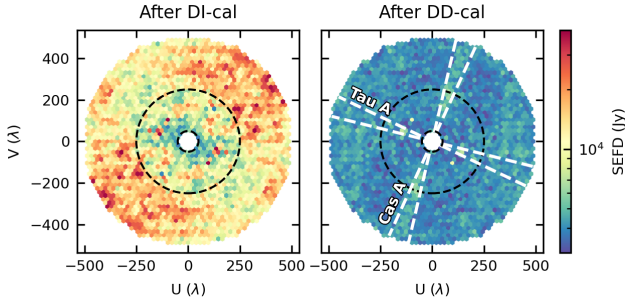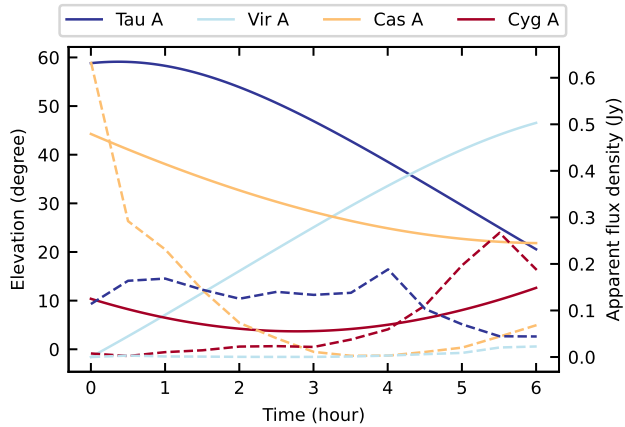
Figure 5. Frequency-integrated (134–147 MHz) dirty images of the 3C 196 field after the DI calibration with only 3C 196 subtracted (top row) and after the DD subtraction (bottom row). The left column shows the all sky images with a pixel scale of 5 arcmin, where a Gaussian taper with FWHM = 10 arcmin and a Briggs weighting with robust parameter −2 have been applied to the *uv*-coverage to better highlight the bright sources. The same colour range has been used in the top and bottom panels to show the level of source sidelobe suppression after the DD subtraction step. The 1, 5, and 50 per cent levels of the time and frequency averaged primary beam are plotted with the dotted contours. The brightest sources are indicated with white circles, including the A-team sources. The right column reports the zoom-in of the 3C 196 central field, imaged with natural weighting and a pixel scale of 0.5 arcmin, covering a 10° × 10° field of view. Only the baselines between 50 and 500λ have been selected. Also here, the same colour range is used in the top and bottom panels. The dashed black circle highlight the 3.9° radius extension of our sky model.

width, which yielded better results than the 1 MHz width for the NCP field (Gan et al. 2023). The gains are solved iteratively for each direction and diagonal elements of the Jones matrix, setting 4 min and 183.1 kHz (i.e. one SB) as solution time and frequency intervals, respectively. This yields enough signal-to-noise for each direction, as discussed in Section 3.4. For the NCP a shorter time interval (i.e.

2.5 min) had to be selected for far-field sources, such as Cas A and Cyg A, because no primary beam was applied and shorter time-scale fluctuations in the beam needed to be solved for. By using an intrinsic sky model and applying the primary beam, we can keep the same time interval for both in and far field sources. Similarly to the NCP data processing, we excluded baselines shorter than 250λ while solv-

Figure 6. Standard deviation of the Stokes I data after DI calibration (left) and DD subtraction (right), expressed in SEFD units, where the SEFD is calculated from the frequency-differenced thermal noise at each uv-cell in the $50-500\lambda$ range. The inner and outer dashed black circles indicate the 50 and 250λ cut used for the power spectrum estimation. The white dashed lines represent the range where we expect the sidelobe contributions of Cas A and Tau A to dominate, given their direction in the sky with respect to the phase centre.

Figure 7. Elevations (left axis, solid lines) and apparent flux densities (right axis, dashed lines) of the brightest A-team sources: Cygnus A (Cyg A), Cassiopeia A (Cas A), Virgo A (Vir A), and Taurus A (Tau A). Line colours correspond to the distance from the phase centre, with dark blue representing the nearest source (Tau A) and dark red the most distant (Cyg A). Apparent flux densities were calculated at 30-minute intervals by multiplying the total flux density of each source by a time and frequency-averaged primary beam model value in the source direction.

ing to avoid over-fitting and signal loss in the uv range of interest for the 21-cm signal extraction, which is $50-250\lambda$ (Mouri Sardarabadi & Koopmans 2019; Mertens et al. 2020; Mevius et al. 2022). The solutions are shown in Fig. B3 for the main field clusters, and in Fig. B4 for the Cas A direction.

After subtracting the sky model with the DD gains applied, a final round of RFI flagging was performed using a base threshold of 1.4, which we found sufficient to remove residual low-level RFI. This resulted in only 0.13 per cent of additional flagged data. We also interpolated the flagged and missing data to reduce the excess noise caused by artificial spectral fluctuations (Offringa et al. 2019a). During this post-DD processing, we flagged baselines shorter than 50λ and longer than 500λ. This uv-cut is usually applied only during the next imaging step (Mertens et al. 2020), but in that case it is

performed on the gridded uv-plane, and hence it depends on the image size. Larger images mean smaller uv-cells, while smaller images mean uv-cells spanning a larger interval of real uv lengths. Cutting at 50λ in the gridded uv-plane would also result in excluding baselines longer than 50λ, biasing the power spectra by setting the image size. A better approach is to directly flag the excluded baselines after DD subtraction and use no uv-cut in the imaging step. Finally, we corrected for the LOFAR element beam. This correction is not necessary in the NCP pipeline because the NCP sky model already has this scaling factor embedded, and the DD gain solutions take it into account. Thus, the power spectrum pipeline was designed to correct only for the array factor. In our case, applying the full primary beam both during gain solving and application kept the visibility data fully apparent, and therefore required the element beam correction to get correct power spectrum values.
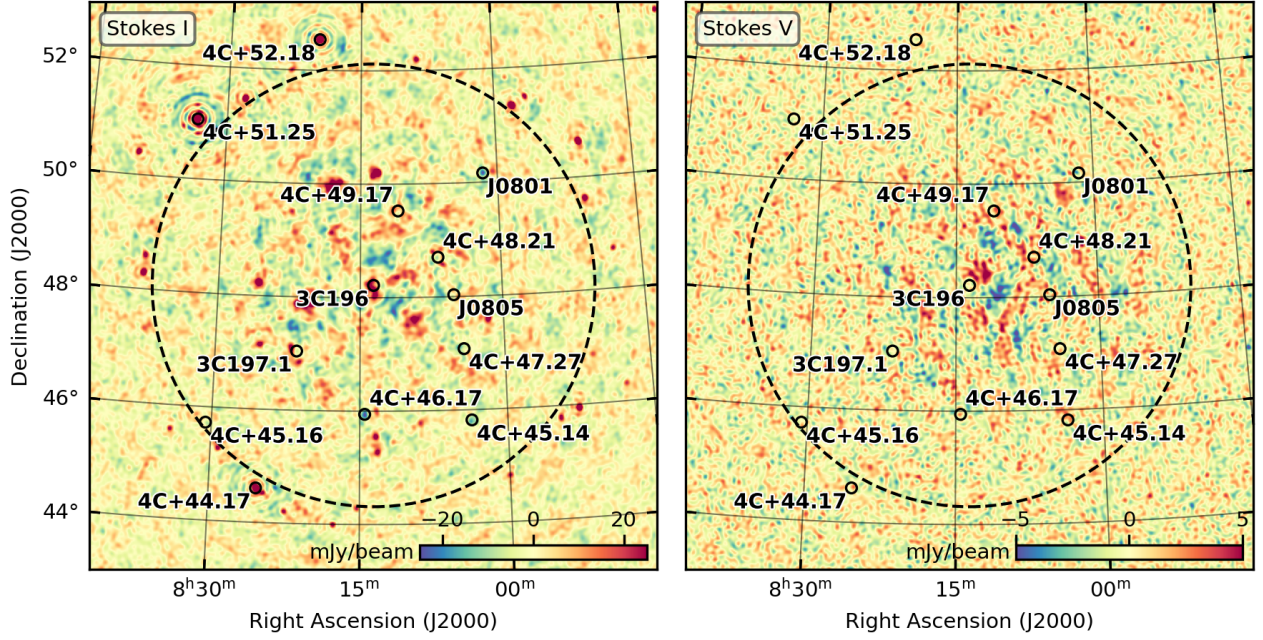
The results after the DD subtraction and final processing are shown in the bottom panels of Fig. 5, where in the all-sky dirty image (bottom left panel) we see that the bright sidelobes from the main field sources have been strongly suppressed and allows us to better distinguish far-field sources such as Tau A. With the same colour scale as for the DI calibrated data (top right panel), the DD subtraction residuals (bottom right panel) show how well the bright sources in the central 3.9° radius (black dashed circle) have been removed. The noise standard deviation of the frequency-integrated images went from $\sigma = 13.6$ mJy/beam of the DI calibrated image to $\sigma = 1.7$ mJy/beam of the residuals after the sky model subtraction. The standard deviation of the uv-gridded Stokes I data after the DD subtraction still shows some residual emission both within and outside 250λ, as shown in the right panel of Fig. 6. However, a reduction of more than one order of magnitude happened, which allows us to see that part of the residual emission comes from Cas A and Tau A, besides a few bright uv-cells where the contamination might come from sources closer to the primary beam main lobe. When the colour scale range of the Stokes I residuals is reduced as shown in Fig. 8, we see that there are a few bright sources not included in the DD sky model but very close to the 3.9° radius cut, the sidelobes of which do affect the central field. Moreover, some extended emission is visible in the Stokes I residuals, which might be un-modelled sky emission or due to calibration errors during the DD solving.

The right panel of Fig. 8 shows the Stokes V image of the SB centred at 140.4 MHz. The Stokes V parameter characterises the circular polarization, which is expected to be close to zero for most of radio sources, and can be used to estimate the noise level. However, we found some deviation from the noise level, especially close to the centre of the field. We do not have a sky model for parameters other than Stokes I, and our calibration did not therefore correct for all the polarization leakage. Jelić et al. (2015) showed that the 3C 196 field exhibits strong polarized structures, which are somewhat similar to what we observed in our Stokes Q and U images (see Appendix C). Leakage from Stokes U into Stokes V is possible, as both are calculated from the XY and YX elements of visibilities. However, we found that the structures in Stokes V more closely resemble those observed in Stokes Q.

4.4 Imaging and conversion to Kelvin

After the DD calibration and final processing steps, the residual visibilities are gridded and imaged using WSCLEAN, creating (l, m, ν) image cubes. To minimize aliasing effects during the gridding, we used a Kaiser-Bessel filter with a kernel size of 15 uv-cells, an oversampling factor of 4096, and 32 w-layers, similar to the settings by Mertens et al. (2020). Offringa et al. (2019b) showed that these set-
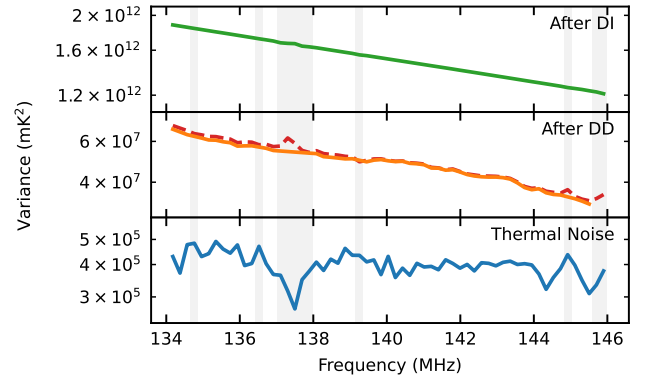
Figure 8. Restored Stokes I (left) and Stokes V (right) images of the SB centred at 140.4 MHz. Both images have a pixel scale of 0.5 arcmin and a field of view of $10° × 10°$. The *uv*-plane was gridded with a natural weighting scheme and only baselines between 50 and $500λ$ were used. The black dashed circle indicates the $3.9°$ radius of the sky model. The brightest sources are highlighted with black circles.

tings keep artefacts introduced by the gridding well below the level of the cosmological 21-cm signal.

The imaging was performed separately for each SB, allowing the production of even and odd time-step Stokes I and V images. The time-differenced Stokes V images will be used to estimate the thermal noise variance. For this reason, having some polarization leakage into Stokes V is not an issue as long as the contamination is not varying rapidly over time. We used a natural weighting scheme, so that every visibility was given a constant weight. We did not set any baseline cut because of the previous baseline flagging, where we selected only the $50−500λ$ range. We set a pixel scale of 30 arcsec and an image size of $512 × 512$ pixels, which covers a field of view of approximately $4.3 × 4.3°$.

To convert the images from units of Jy/beam to units of brightness temperature (Kelvin), we Fourier transformed the $(l, m, ν)$ image cube into a gridded $(u, v, ν)$ visibility cube. The conversion was then performed using the method outlined by Offringa et al. (2019b). A spatial Tukey taper with a $4°$ diameter was also applied to focus on the central part of the primary beam, which has FWHM $≈ 4.1°$ at 140 MHz, and to avoid sharp image edges.

Before proceeding to power spectrum estimation, we performed additional flagging to remove any outliers in the gridded data cubes, using a k-sigma clipping method with de-trending. These outliers are mainly low-level RFI that AOFLAGGER failed to detect. In particular, *uv*-grid outliers are flagged based on their weights (i.e. the per-visibility inverse variance combined with the weights given by the *uv*-density of the gridded visibilities), Stokes V variance, standard deviation of time-differenced Stokes V, and standard deviation of frequency-differenced Stokes I (i.e. right panel of Fig. 6). Frequency outliers are flagged based on their weights and standard deviation of frequency-differenced Stokes V. Figure 9 shows the variance in mK^2 over frequency for the Stokes I residual cubes before (red dashed line) and after (orange solid line) this last flagging. The shaded grey

Figure 9. Variance in mK^2 units as a function of frequency for the Stokes I after DI calibration (green solid line), after DD subtraction (red dashed line), and after final outlier flagging (orange solid line, see Section 4.4). The thermal noise estimated from the time-differenced Stokes V visibilities is shown with the blue solid line. The gray shaded areas represent the flagged channels.

areas represent the resulting flagged channels. We found that 6.4 per cent of *uv*-cells and 16.4 per cent of the SB are flagged, which are percentages lower than the values found in the NCP processing (Mertens et al. 2020).

4.5 Power spectrum estimation

Let $T(\mathbf{x})$ represent the brightness temperature of a signal at a physical coordinate $\mathbf{x}$. The corresponding power spectrum, denoted as $P(\mathbf{k})$, is a function of the wavenumber $\mathbf{k}$ (with units of $h\,\mathrm{cMpc}^{-1}$) and can

be expressed as

$$P(\mathbf{k}) = V|\tilde{T}(\mathbf{k})|^2 \,, \tag{2}$$

where V is the observing comoving volume, and $\tilde{T}$ is the discrete Fourier transform of the brightness temperature. This power spectrum is typically reported in units of $\mathrm{K}^2 \, h^{-3} \, \mathrm{cMpc}^3$. The components of $\mathbf{k}$, perpendicular and parallel to the line of sight, are respectively given by (Morales & Hewitt 2004; Vedantham et al. 2012):

$$\mathbf{k}_\perp = \frac{2\pi}{D_M(z)}\mathbf{u}, \; k_\parallel = \frac{2\pi H_0 \nu_{21} E(z)}{c(1+z)^2}\,\eta\,, \tag{3}$$

where $D_M(z)$ is the transverse comoving distance at redshift z, $\nu_{21} = 1420\,\mathrm{MHz}$ is the rest frequency of the neutral hydrogen hyperfine transition line, H_0 is the Hubble constant, $E(z)$ represents the dimensionless Hubble parameter, c is the speed of light, and η is the Fourier dual of the frequency ν.

The cylindrically averaged (2D) power spectrum can be obtained by averaging $P(\mathbf{k})$ in cylindrical shells with radius $k_\perp = |\mathbf{k}_\perp|$:

$$P(k_\perp, k_\parallel) = \frac{\sum_{\mathbf{k}\in(k_\perp,k_\parallel)} P(\mathbf{k})}{N_{(k_\perp,k_\parallel)}}\,, \tag{4}$$

where $N_{(k_\perp,k_\parallel)}$ is the number of $\mathbf{k}$-space cells within the $(k_\perp, k_\parallel)$-annulus. Alternatively, the dimensionless spherically averaged (1D) power spectrum can be derived by averaging in spherical shells with radius $k = |\mathbf{k}| = \sqrt{k_\perp^2 + k_\parallel^2}$:

$$\Delta^2(k) = \frac{k^3}{2\pi^2}\frac{\sum_{\mathbf{k}\in k} P(\mathbf{k})}{N_k}\,, \tag{5}$$

where N_k is the number of $\mathbf{k}$-space cells within the k-shell. In 21-cm cosmology, the spherical power spectrum is in units of K^2, and because of the direct connection to brightness temperature units, the 21-cm signal upper limits are typically reported as $\Delta^2(k)$. In Equations (4) and (5), $P(k_\perp, k_\parallel)$ and $\Delta^2(k)$ are optimally weighted according to the gridded visibility thermal noise (more details in Mertens et al. 2020).

The wavenumber $k_\parallel$ is effectively the Fourier conjugate of frequency, and spectrally smooth foregrounds should remain confined at low $k_\parallel$ values, whereas the 21-cm signal affects a wider range of $k_\parallel$ modes (e.g. Santos et al. 2005). In addition to this mode separation, we have to consider the intrinsic chromaticity of interferometers, which spreads foreground power at higher $k_\parallel$, causing a process called 'mode-mixing' (Morales et al. 2012, 2019). Because of the characteristic shape that this effect assumes in the cylindrically averaged power spectrum, it is also denoted as the 'foreground wedge' (Datta et al. 2010; Vedantham et al. 2012; Liu et al. 2014a,b). The extent of the foreground wedge is defined by the horizon delay line, which can be derived under a formalism accounting for curved sky effects following Munshi et al. (2025b) for phase-tracking instruments such as LOFAR. The same formalism also prescribes source lines which specify a range in the cylindrical power spectrum where emission from a specific direction in the sky is expected to cause power. Because all the sky emission stays within the wedge under reasonable assumptions, above the horizon line there is a foreground-free region called 'EoR window', where ideally the power spectrum should be consistent with the noise power spectrum for the sensitivity of the current generation of interferometers.

The power spectra presented in this paper were estimated using the power spectrum pipeline PSPIPE[14]. We selected the $50-250\lambda$ baseline

range, filtering out the short and long baselines. The reported power spectrum uncertainties were estimated from the sample variance, as described by Mertens et al. (2020).

In Fig. 10, we show the cylindrical power spectra of Stokes I, V, and thermal noise (i.e. time-differenced Stokes V) before (top row) and after (bottom row) the outlier flagging discussed in Section 4.4. We applied a Blackman-Harris window function to mitigate the effect of the limited frequency bandwidth leaking power into the EoR window. The power spectra before the flagging show a peak of power at $k_\perp \approx 0.14\,h\,\mathrm{cMpc}^{-1}$, corresponding to a gap in the uv-coverage which has been observed also in NCP data (Mertens et al. 2020). Flagging the outliers in the uv-plane lowered the variance, making the thermal noise more uniform and the residual foreground confined at low $k_\parallel$. As seen in Fig. 8, the power spectrum of Stokes V shows residual emission at low $k_\parallel$ because of polarization leakage. However, this leakage is time-correlated and is not visible in the thermal noise, making this effect not a concern for the current analysis. Similar to Fig. 6, the cylindrical power spectra also show some residual emission in the direction of Cas A (dashed black lines), which is mainly visible in Stokes V and at low $k_\perp$ in Stokes I. The outlier flagging removed most of the affected uv-cells, leaving some residuals only at $k_\perp \lesssim 0.08\,h\,\mathrm{cMpc}^{-1}$.

5 RESIDUAL FOREGROUND REMOVAL

As shown in Fig. 10, the cylindrical power spectra are still dominated by residual foreground contamination, especially at low $k_\parallel$. These modes are important because it is where the 21-cm signal is expected to be stronger. These residuals are dominated by diffuse Galactic emission, sources near or below the confusion noise, and sources outside the $3.9°$ radius that are not included in the DD sky model. The distinction between foregrounds and the 21-cm signal lies in their spectral behaviour: foregrounds exhibit a much larger coherence scale than both thermal noise and the 21-cm signal. By subtracting smoothed solutions during the DD calibration step, we made sure that the residual foreground contamination also remained smooth. This property can be used to isolate and subtract the residual foregrounds. In this work, we used the GPR method described by Mertens et al. (2018) to model these components. The model basis of the 21-cm signal has been built from simulations by using a machine learning (ML) training, which resulted in an enhanced GPR method called ML-GPR (Mertens et al. 2024).
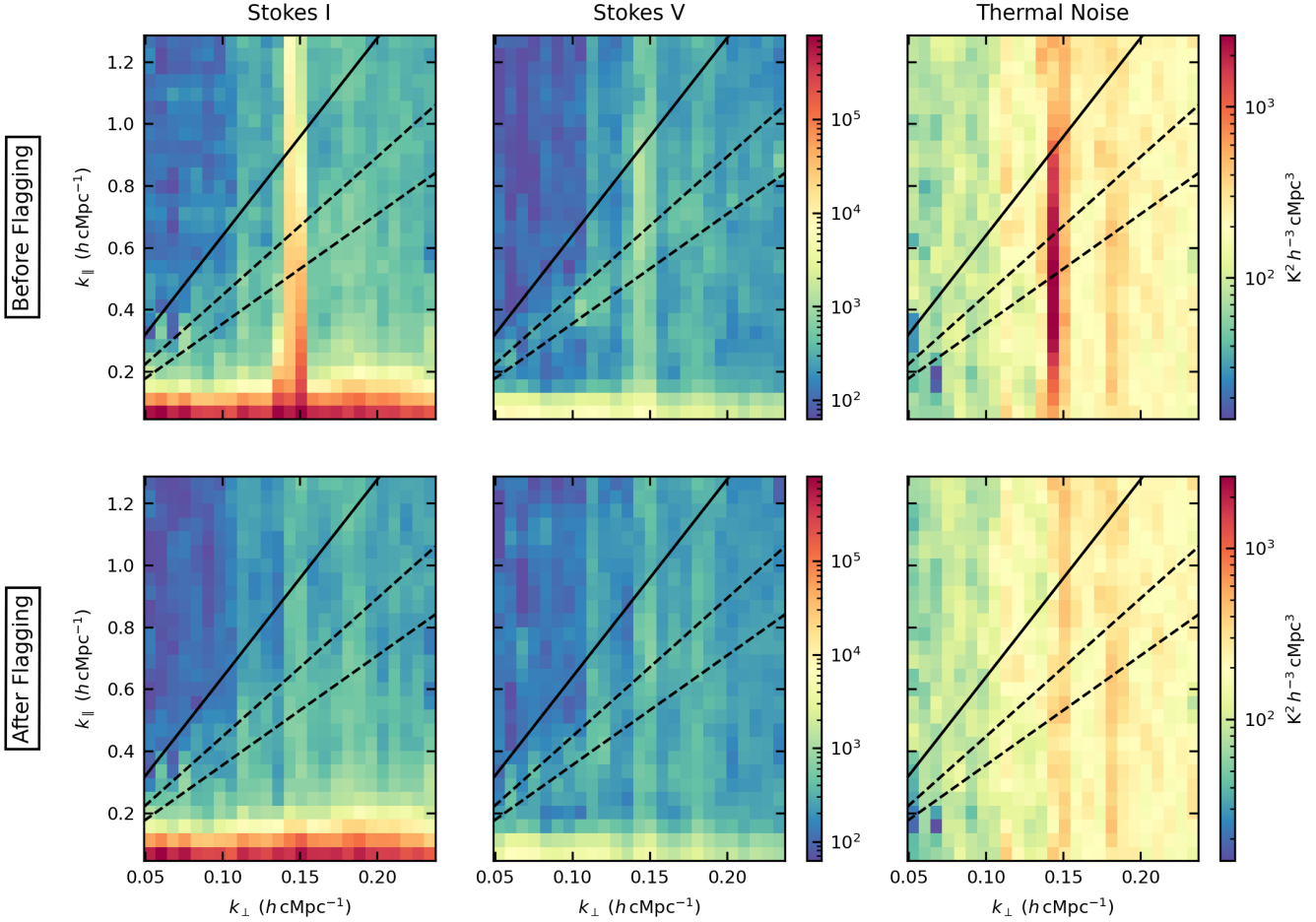
5.1 Gaussian process regression

Gaussian process regression is a non-parametric Bayesian method used to model and predict data based on prior knowledge, where the data is assumed to follow a Gaussian process. In the context of 21-cm cosmology, GPR models the observed data as a sum of Gaussian processes that represent the foregrounds, thermal noise, and the 21-cm signal, each characterised by a specific frequency covariance function κ (also called kernel) and zero mean. These functions are described by adjustable hyper-parameters that define properties such as the variance, coherence scale, and shape. These hyper-parameters are optimised by maximising their posterior probability based on the observed data. Once the optimal model is obtained, the expected values of the foreground components are subtracted from the data.

Following the standard GPR framework described by Mertens et al. (2018), the observed data $\mathbf{d}$ at frequencies ν are modelled by foreground $\mathbf{f}_{\mathrm{fg}}$, 21-cm signal $\mathbf{f}_{21}$, and noise $\mathbf{n}$ components:

$$\mathbf{d}(\nu) = \mathbf{f}_{\mathrm{fg}}(\nu) + \mathbf{f}_{21}(\nu) + \mathbf{n}(\nu)\,. \tag{6}$$

Figure 10. Cylindrical power spectra of Stokes I (first column), Stokes V (second column), and thermal noise (third column). The top row shows the power spectra before the outlier flagging discussed in Section 4.4, while the bottom row shows them after such a flagging. In all panels, the solid line indicates the horizon delay line, whereas the dashed lines indicate the delay ranges where we expect most of the power of Cas A. The same colour bar is used for the panels of the same column, and for Stokes I and Stokes V power spectra.

In our case, the **d** is the gridded (u, v, ν) visibility cube before making the power spectrum. This approach in the uv-space allows GPR to account for the baseline dependence of the frequency coherence scales, effectively modelling both the foreground wedge and thermal noise. By exploiting the distinct spectral behaviours of the different components, we can model the total covariance matrix as a sum of the individual GP covariance matrices:

$$\mathbf{K}(\nu, \nu) = \mathbf{K}_{\mathrm{fg}}(\nu, \nu) + \mathbf{K}_{21}(\nu, \nu) + \mathbf{K}_{\mathrm{n}}(\nu, \nu), \qquad (7)$$

where $\mathbf{K}_{\mathrm{fg}}$ represents the smooth foregrounds, $\mathbf{K}_{21}$ the 21-cm signal, and $\mathbf{K}_{\mathrm{n}}$ the noise. The element (p, q) of the covariance matrix corresponds to $\kappa(\nu_p, \nu_q)$, which is defined between two points ν_p and ν_q, which, in our case, are two frequency channels (i.e. SB). Because **n** is a Gaussian-distributed, frequency-uncorrelated noise with variance σ_{n}^2, the noise covariance matrix is $\mathbf{K}_{\mathrm{n}}(\nu, \nu) = \mathrm{diag}[\sigma_{\mathrm{n}}^2(\nu)]$.

The foreground covariance includes two components: an intrinsic foreground $\mathbf{K}_{\mathrm{int}}$ to capture the large frequency coherence scale of extragalactic and Galactic emissions, and a mode-mixing $\mathbf{K}_{\mathrm{mix}}$ for the smaller frequency coherence scale (1−5 MHz) of the foreground wedge. While the instrument chromaticity causes the mode-mixing to be a multiplicative effect, it can still be approximated as additive

to first order, allowing us to define $\mathbf{K}_{\mathrm{fg}} = \mathbf{K}_{\mathrm{int}} + \mathbf{K}_{\mathrm{mix}}$, similar to Mertens et al. (2020).

As we will show in Section 5.2, the data can not be fully described by just the foreground and the 21-cm signal components. Similar to the NCP analysis (Mertens et al. 2020; Munshi et al. 2024), also for the 3C 196 data we observed an additional source of power with a small frequency coherence scale that is difficult to distinguish from the 21-cm signal. This 'excess power' may arise from various instrumental effects, low-level RFI, polarization leakage, or calibration errors. We had to introduce an additional covariance matrix $\mathbf{K}_{\mathrm{ex}}$ to capture the additional complexity of the data.

We can then define our total Gaussian process as the joint probability density distribution of the data **d** and function values $\mathbf{f}_{\mathrm{fg}}$ of the foreground model at frequencies ν:

$$\begin{bmatrix} \mathbf{d} \\ \mathbf{f}_{\mathrm{fg}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\mathrm{fg}} + \mathbf{K}_{21} + \mathbf{K}_{\mathrm{ex}} + \mathbf{K}_{\mathrm{n}} & \mathbf{K}_{\mathrm{fg}} \\ \mathbf{K}_{\mathrm{fg}} & \mathbf{K}_{\mathrm{fg}} \end{bmatrix} \right), \qquad (8)$$

where we used the short-hand $\mathbf{K} \equiv \mathbf{K}(\nu, \nu)$. Because our data do not currently have the sensitivity for a detection of the 21-cm signal and our knowledge of the excess power is limited, we kept a conservative approach by subtracting only the modelled foreground components

from the observed data:

$$\mathbf{d}_{\mathrm{res}} = \mathbf{d} - \mathbb{E}[\mathbf{f}_{\mathrm{fg}}] \,, \tag{9}$$

where $\mathbb{E}[\mathbf{f}_{\mathrm{fg}}]$ is the expectation value of the foreground model, given by

$$\mathbb{E}[\mathbf{f}_{\mathrm{fg}}] = \mathbf{K}_{\mathrm{fg}}[\mathbf{K}_{\mathrm{fg}} + \mathbf{K}_{21} + \mathbf{K}_{\mathrm{ex}} + \mathbf{K}_{\mathrm{n}}]^{-1}\mathbf{d} \,. \tag{10}$$

The covariance of the foreground model is defined as

$$\mathrm{cov}[\mathbf{f}_{\mathrm{fg}}] = \mathbf{K}_{\mathrm{fg}} - \mathbf{K}_{\mathrm{fg}}[\mathbf{K}_{\mathrm{fg}} + \mathbf{K}_{21} + \mathbf{K}_{\mathrm{ex}} + \mathbf{K}_{\mathrm{n}}]^{-1}\mathbf{K}_{\mathrm{fg}} \,. \tag{11}$$

Equations (10) and (11) can be generalised to any component $\mathbf{f}$ of our model to estimate its power spectrum and uncertainty. To achieve this, we draw m samples from the posterior distribution of the hyper-parameters. For each sample, we compute $\mathbb{E}[\mathbf{f}]$ and $\mathrm{cov}[\mathbf{f}]$ using the aforementioned equations. A power spectrum is then calculated by adding to $\mathbb{E}[\mathbf{f}]$ a sample drawn from a Gaussian distribution with covariance $\mathrm{cov}[\mathbf{f}]$. The final power spectrum and its 1σ uncertainty are determined as the median and standard deviation of the m power spectra at each k-mode.

5.2 Covariance model

Because the GPR model has zero mean, the components are completely defined by their covariance functions. Describing their correct form and shape is then fundamental. To model the foregrounds and the excess power, we found that the class of Matern covariance functions describes both smooth and rough frequency variations well. Its analytical form is given by (Stein 1999)

$$\kappa_{\mathrm{Matern}}(\nu_p, \nu_q) = \sigma^2 \frac{2^{1-\mu}}{\Gamma(\mu)} \left(\frac{\sqrt{2\mu}\, r}{l}\right)^{\mu} K_{\mu}\left(\frac{\sqrt{2\mu}\, r}{l}\right) \,, \tag{12}$$

where σ^2 is the variance, $r = |\nu_q - \nu_p|$ is the absolute difference between the two frequencies, l is the frequency coherence scale, μ is the smoothness parameter, Γ is the Gamma function, and K_{μ} is the modified Bessel function of the second kind. Throughout this paper, all the σ^2 values are expressed as a fraction of the variance of the observed data. We used different kernels for the 21-cm signal and the thermal noise. Below, we describe the different covariance kernels used in our GPR model:

Intrinsic foreground – $\mathbf{K}_{\mathrm{int}}(\sigma_{\mathrm{int}}^2)$: The intrinsic foreground is constituted by the residual extragalactic and Galactic emission within the field of view. Because it is expected to be spectrally smooth, we modelled the covariance function with a radial basis function (RBF), which is a Gaussian covariance function obtained by setting $\mu = \infty$ (Mertens et al. 2018). It is the Matern function with the quickest drop at high $k_{\parallel}$, so it is ideal to model the residual foregrounds within the primary beam that are confined at low $k_{\parallel}$, as shown in Fig. 10. The function is characterised by two main hyper-parameters, the frequency coherence scale l_{int} and the variance σ_{int}^2. Initially, we set a uniform prior $\mathcal{U}(10, 100)$ MHz on l_{int}, but we found it was consistently unconstrained with a lower limit of ≈ 60 MHz after the optimisation. This indicated that l_{int} was significantly larger than the data bandwidth (i.e. 12 MHz), and the model was unable to recover the effective coherence scale. We therefore fixed l_{int} to 80 MHz, to reduce the number of degrees of freedom and speed up the fitting process. This choice did not affect the estimated values of the other hyper-parameters.

Mode-mixing foreground – $\mathbf{K}_{\mathrm{mix}}(\theta_{\mathrm{mix}}, \sigma_{\mathrm{mix}}^2)$: Mode-mixing, caused by the chromatic response of the instrument, introduces smaller frequency coherence scales, typically in the range of 1–5 MHz. We found that using a Matern covariance function with $\mu = 3/2$ provided the largest marginal likelihood (i.e. evidence) for the model. This is the same shape used by Mertens et al. (2020), where the two hyper-parameters l_{mix} and σ_{mix}^2 were optimised. The result was a kernel with no dependence on baseline length, whereas mode-mixing effects should produce a wedge-like structure in the cylindrical power spectrum that depends on baseline length (Datta et al. 2010; Morales et al. 2012). To take into account this $k_{\perp}$ dependence of the scale length, we modified the Matern covariance function to parametrize such a wedge feature into the frequency coherence scale:

$$l_{\mathrm{mix}}(u) = \frac{\nu_z}{\nu_z\, \eta_{\mathrm{buffer,mix}} + u \sin\theta_{\mathrm{mix}}} \,, \tag{13}$$

where $\eta_{\mathrm{buffer,mix}} \propto k_{\parallel}$ is the delay buffer, $u \propto k_{\perp}$ is the baseline length (see Equation 3), θ_{mix} is the angle of the wedge-like structure in radians, and ν_z is the mean frequency of the redshift bin z. The delay buffer is used to ensure that the wedge does not start from $k_{\parallel} = 0$, adding a sort of extra intrinsic foreground. We found that $\eta_{\mathrm{buffer,mix}}$ converged to 0.1 µs for our mode-mixing component. To decrease the number of degrees of freedom, we fixed this hyper-parameter to that value, leaving only θ_{mix} and σ_{mix} to be optimised. We set a uniform prior $\mathcal{U}(0.01, 1.6)$ rad on θ_{mix}, where the upper bound is $\approx \pi/2$ that is the maximum angle allowed for the horizon line. An issue with such a wedge parametrization is that Equation (13) assumes the flat-sky horizon lines, but Munshi et al. (2025b) showed that the angle can be very different for phasing arrays such as LOFAR. An implementation of this improved wedge parametrization is left for the future.

Excess power – $\mathbf{K}_{\mathrm{ex}}(l_{\mathrm{ex}}, \alpha_{\mathrm{ex}}, \sigma_{\mathrm{ex}}^2)$: Most of our GPR efforts were put on finding the best kernel to describe the excess power. Even though recent works pointed out that such an excess is in large part related to DD gain errors on bright distant sources (Gan et al. 2022; Brackenhoff et al. 2024, 2025; Ceccotti et al. 2025), we do not know for certain its cause. Therefore, no prior knowledge was used to set up the excess power kernel shape. We found that the excess in our 3C 196 data dominates the lower k-modes, in a 'mode-mixing foreground'-like behaviour. While keeping all the other components unchanged and the same covariance function shape (we started with $\mu = \infty$, i.e. an RBF kernel), we saw that by using the wedge parametrization the marginal likelihood of the GPR model increased by a few per cent. However, for the excess, θ was giving $\pi/2$ as a lower limit. Because that is the maximum angle allowed by Equation (13), and we know that the real horizon limit has an angle larger than the flat-sky horizon line, we had to test a different parametrization to push this limit. Instead of using an angle θ, we can express the $k_{\perp}$ dependence of the coherence scale and the variance as

$$l_{\mathrm{ex}}(u) = \frac{l_{0,\mathrm{ex}}}{1 + 10^{-3}\alpha_{\mathrm{ex}}(u - u_{\mathrm{min}})} \,, \quad \sigma_{\mathrm{ex}}^2(u) = \sigma_{0,\mathrm{ex}}^2\left(\frac{u}{u_{\mathrm{min}}}\right)^{\alpha_{\sigma,\mathrm{ex}}} \,, \tag{14}$$

where $l_{0,\mathrm{ex}}$ is the coherence scale at $u = u_{\mathrm{min}}$, $u_{\mathrm{min}} = 50\lambda$ is the minimum baseline length, and α_{ex} and $\alpha_{\sigma,\mathrm{ex}}$ are the coefficients that encode the baseline dependence into the coherence scale and the variance, respectively. Given the lack of prior knowledge on the excess kernel, we set a uniform prior $\mathcal{U}(-5, 60)$ on α_{ex}, where the large upper bound was to allow a steep wedge (even above the expected real horizon limit), while the lower negative bound was to rule out an inverse wedge-like structure. For $l_{0,\mathrm{ex}}$, we set a uniform

prior $\mathcal{U}(0.2, 0.8)$ MHz, which gives a very small coherence scale. Keeping the same prior ranges, we found that using a Matern covariance function with $\mu = 5/2$ gave higher marginal likelihood than an RBF and $\mu = 3/2$. While we would expect a flat baseline dependence for the variance, the data preferred $\alpha_{\sigma,\mathrm{ex}} < 0$. To decrease the degrees of freedom, we fixed this parameter to its converging value, namely $\alpha_{\sigma,\mathrm{ex}} = -0.25$. We report some of the most relevant tests in Appendix D.

21-cm signal – $\mathbf{K}_{21}(x_1, x_2, \sigma_{21}^2)$. Instead of using a Matern covariance function, the 21-cm signal kernel was constructed using an ML-based variational auto-encoder (VAE) kernel, as described by Mertens et al. (2024) and Acharya et al. (2024). The VAE kernel is trained on simulations to compress the high-dimensional 21-cm signal data into a lower-dimensional latent space, from which the covariance matrix is reconstructed. This approach was necessary because standard GPR kernels could struggle to capture the complex frequency covariance of the 21-cm signal in the observed data, leading to biases and risk of signal loss in the power spectrum estimation (Kern & Liu 2021). The VAE kernel, in contrast, adapts to the different physical characteristics of the signal learned from 21-cm signal simulations. This ML-GPR approach has already been applied by Munshi et al. (2024).

For this work, the VAE kernel was trained on 21-cm signal models for $z = 9.16$, using the simulation framework described by Acharya et al. (2024). The EoR simulations were generated with GRIZZLY (Ghara et al. 2015a,b, 2018, 2020), a 1D radiative transfer code coupled with cosmological and N-body simulations to produce 21-cm brightness temperature maps at different redshifts. The ML-trained kernel resulted in a covariance function described by only two latent space dimensions, x_1 and x_2, and a scaling factor for the 21-cm signal variance σ_{21}^2. During the training, the VAE forces the latent space to be normally distributed with variance one. We set a uniform prior $\mathcal{U}(-3, 3)$ on x_1 and x_2, a choice that is sufficient to explore any possible expected shape of the 21-cm signal power spectra.

Thermal noise – $\mathbf{K}_n$: The noise covariance function is built from the time-differenced Stokes V visibilities, which can be slightly lower than the noise in Stokes I. Therefore, we estimated the scaling factor using an initial uniform prior $\mathcal{U}(0.5, 2.5)$ and consistently found a value of one across different tests. We then fixed the scaling factor to this value to reduce the number of degrees of freedom and reduce the computing time.

5.3 Application to data and residual power spectrum

The input data for our ML-GPR model consists of the gridded visibility cubes after the outlier flagging, as described in Section 4.4. The posterior probability distributions and the Bayesian evidence of the covariance model were derived with the ULTRANEST[15] package (Buchner 2021), based on a nested sampling Monte Carlo algorithm (Buchner 2016, 2019). We used 100 live points to explore the parameter space within the prior constraints and find the posterior distribution of the hyper-parameters.

The prior ranges and the results of the parameter estimation are presented in Table 2, with a corner plot of the posterior distributions shown in Fig. 11. The posterior distributions of the foregrounds and excess parameters are peaked and symmetric around the estimated values, giving small uncertainties. Parameters of the same component

---

Figure 11. Posterior probability distribution for each of the ML-GPR hyperparameters (Table 2). The contours represent the 68, 95, and 99.7 per cent confidence intervals. The vertical dashed lines in the histograms indicate the 16th and the 84th percentile of the distributions.

show low to moderate correlation, with a high negative correlation between σ_{mix}^2 and θ_{mix}. Because the 21-cm signal component is well below the thermal noise level, its parameters x_1 and x_2 did not converge and give all the allowed signal shapes equally probable, as should be expected. For the same reason, its variance σ_{21}^2 reached an upper limit.

From these posterior distributions, we sampled 500 realisations of each component cube. For each realisation, we estimated a power spectrum and we took the median of these 500 spectra at each k-mode. The resulting cylindrical power spectra $P(k_\perp, k_\parallel)$ are shown in Fig. 12. While intrinsic and mode-mixing foregrounds are similar to the NCP results, the excess component for our 3C 196 data shows a foreground-like feature, in contrast to a more noise-like behaviour for the NCP excess (Mertens et al. 2020). Our excess is characterised by a wedge structure confined within the real horizon limit (solid black line), given by $\alpha_{\mathrm{ex}} = 36 \pm 1$. The peak in power observed in the bottom-left corner of the $k_\perp$-$k_\parallel$ space suggests that the source of such an excess may be related to residual extended emission outside the primary beam. A few bright $(k_\perp, k_\parallel)$-cells are located along the Cas A direction (dashed black lines), where the DD-subtracted data showed some residual emission (see Fig. 6). The kernel large extent in $k_\parallel$ causes some leakage into the EoR window, where the power is approximately an order of magnitude higher than the modelled mode-mixing foreground. However, as we will discuss later, the contamination in the EoR window remains at or below the thermal noise level. The residual data cube was then obtained by subtracting the sampled realisations of the two foreground components from the input data (Equation 9).

A summary of the power spectra of input data, ML-GPR components, and residual data is shown in Fig. 13, where the $k_\perp$-averaged power spectrum $P(k_\parallel)$ and the spherical power spectrum $\Delta^2(k)$ are shown in the top and bottom panel, respectively. The excess component is higher than the noise at $k \approx k_\parallel \lesssim 0.3\,h\,\mathrm{cMpc}^{-1}$, but goes
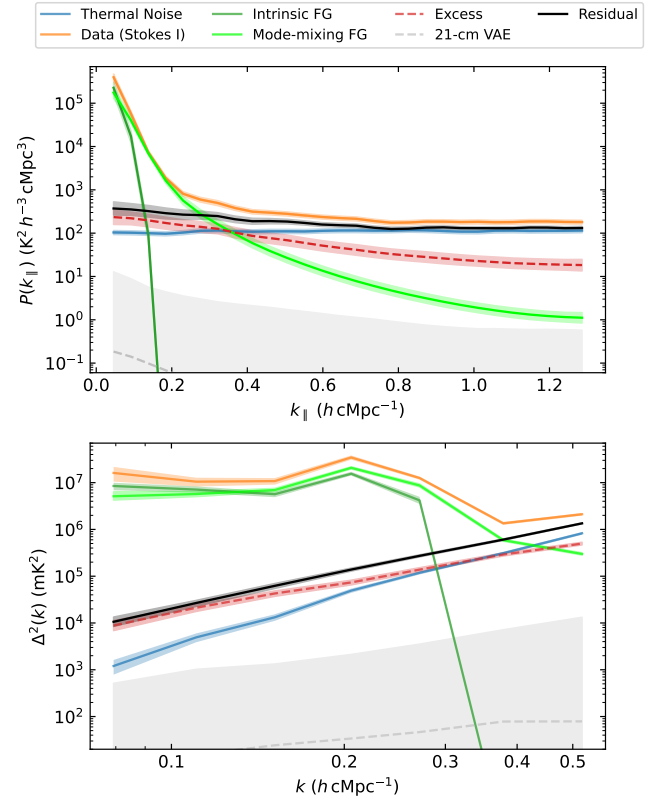
Table 2. Components of our ML-GPR model, along with the parameter priors and the estimated median values with the 1σ uncertainties. All the variances are expressed as a fraction of the input data.

Component	Covariance	Parameter	Description	Prior	Estimated value
Intrinsic foregrounds	Radial Basis Function	σ^2_{int}	Variance	$\log \mathcal{U}(-0.5, 0.5)$	-0.38 ± 0.02
Mode-mixing foregrounds	Matern Function ($\eta = 3/2$)	σ^2_{mix}	Variance	$\log \mathcal{U}(-0.8, -0.01)$	-0.34 ± 0.01
		θ_{mix}	Angle (rad)	$\mathcal{U}(0.01, 1.6)$	0.175 ± 0.003
21-cm signal	Trained VAE-based Kernel	x_1	Latent space dimension	$\mathcal{U}(-3, 3)$	–
		x_2	Latent space dimension	$\mathcal{U}(-3, 3)$	–
		σ^2_{21}	Variance	$\log \mathcal{U}(-7, -0.5)$	< -3.2
Excess power	Matern Function ($\eta = 5/2$)	$\sigma^2_{0,\text{ex}}$	Variance	$\log \mathcal{U}(-4, -2)$	-2.51 ± 0.2
		$l_{0,\text{ex}}$	Length scale (MHz)	$\mathcal{U}(0.2, 0.8)$	0.52 ± 0.02
		α_{ex}	Baseline dependence	$\mathcal{U}(-5, 60)$	36 ± 1

Figure 12. Cylindrical power spectra of the components of our ML-GPR model: intrinsic foreground (top left), mode-mixing component (top right), excess power (bottom left), and 21-cm signal (bottom right). The top row shows the components that are subtracted from the observed data, whereas the bottom row shows the components that will form the residual data along with the thermal noise. In all panels, the solid line indicates the horizon delay line, whereas the dashed lines indicate the delay ranges where we expect most of the power of Cas A. The same colour range is used for the intrinsic and the mode-mixing foregrounds.

Figure 13. Decomposition of the components of our ML-GPR model for the input residual Stokes I data (orange solid line). The covariance model is constituted by intrinsic foreground (green solid line), mode-mixing component (lime solid line), excess (red dashed line), 21-cm signal (gray dashed line), and thermal noise (blue line). The ML-GPR residual data (black solid line) are obtained by subtracting the foreground components from the input data. The 2σ uncertainties are reported with the shaded areas. The top panel shows the cylindrical power spectra as a function of $k_\parallel$ (i.e. averaged over $k_\perp$), whereas the bottom panel shows the spherical power spectra.

below it at higher k, becoming a fraction of the noise power. This behaviour resembles the power spectrum of a foreground component but with much smaller spectral coherence. In these plots we also show the 21-cm signal component, whose variance was at least more than two orders of magnitude lower than the data variance and the parameters x_1 and x_2 were unconstrained. This results in a genuine upper limit on the 21-cm signal power spectrum with albeit large uncertainties, consistent with Mertens et al. (2024).

Figure 13 also shows the residual power spectra. Because only the foreground components were subtracted from the input data, the residual power spectrum mainly consists of the excess at low $k_\parallel$ and noise at high $k_\parallel$. The ratio between the cylindrical power

spectra of residual and thermal noise is shown in Fig. 14. The EoR window is very clean, with a mean ratio of approximately 1.2. On the other hand, the ratio is approximately 3.0 on average within the wedge, being dominated by the excess component especially at $k_\perp <$ 0.11 h cMpc^{-1}, where the mean ratio is approximately 5.8. Note that the noise power spectrum estimated from time-differenced Stokes V

Figure 14. Cylindrical power spectrum of the ML-GPR residuals divided by the power spectrum of the thermal noise (i.e. time-differenced Stokes V). The grey solid line indicates the horizon delay line, whereas the grey dashed lines indicate the delay ranges where we expect most of the power of Cas A. The colour range between 1 and 14 is chosen to allow an easier comparison with Fig. 12 in Mertens et al. (2020) .

is just a realisation drawn from the theoretical noise distribution, which is expected to have a constant power at fixed $k_\perp$. This means that every $k_\parallel$-cell at fixed $k_\perp$ should have a value more or less close to the mean. However, there are a few $(k_\perp, k_\parallel)$-cells where the observed thermal noise is more than 2σ lower or higher than the mean value, resulting in a quite high or low ratio, respectively. This is the case for the very high ratios at $k_\perp \lesssim 0.07\,h\,\mathrm{cMpc}^{-1}$ and $k_\parallel \approx 0.15\,h\,\mathrm{cMpc}^{-1}$.

6 VALIDATION OF THE PROCESSING PIPELINE

Because the foreground emission is several orders of magnitude brighter than the 21-cm signal, the data calibration and foreground removal steps might suppress or bias the 21-cm results. While the power spectrum estimation pipeline has been extensively tested against known power spectra (Mertens et al. 2020), the calibration steps and the ML-GPR foreground removal need to be validated against signal loss.

6.1 Robustness test on calibration steps

The main danger of the DI-calibration (Section 4.2) is setting an incorrect flux scale, biasing the final results by a scaling factor. The bottom panel of Fig. 2 shows that the peak brightness of 3C 196 after the DI-steps (black dots) agrees with the expected total flux of the source (Scaife & Heald 2012). This ensures that the flux scale is correctly set for the DI-calibrated data. The DD solving and sky model subtraction (Section 4.3) could affect the resulting power spectra more drastically because they have the potential to modify and remove part of the 21-cm signal (Patil et al. 2016; Ewall-Wice et al. 2017). This bias is strongly curtailed by solving gains

using only baselines longer than 250λ, as shown by Mevius et al. (2022). Moreover, we applied a smoothing kernel of 4 MHz width, such that over-fitting and 21-cm signal suppression are reduced. This smoothing is different from the standard method employed in the NCP processing (Patil et al. 2017; Mertens et al. 2020, 2025), which makes use of sagecal (Yatawatta 2016). Instead of smoothing the solution with a Gaussian kernel at each iteration, as done in ddecal, sagecal uses a consensus optimisation algorithm to spectrally constrain the solutions with a third-order Bernstein polynomial (more details in Yatawatta 2015, 2016; Yatawatta et al. 2017; Yatawatta 2018). Gan et al. (2023) showed that ddecal with our settings gives results on the 21-cm power spectrum comparable to sagecal, implying limited signal loss in the DD calibration step.

6.2 Robustness test on ML-GPR

The GPR signal-separation method has been extensively tested, and its reliability in statistically separating foreground and 21-cm signal was confirmed (Mertens et al. 2018; Offringa et al. 2019a). However, any small error in the foreground components might completely alter the inferred 21-cm signal because of their up to five orders of magnitude difference in power. The new ML-GPR method limits this problem by using an ML-based VAE kernel, but a bias in the excess component might still happen.

The validation process of the ML-GPR foreground removal is similar to that by Mertens et al. (2020), where synthetic 21-cm signals are generated and added to the data. We performed such a signal injection test with the following steps:

(i) *Signal generation*: The trained 21-cm VAE kernel was used to generate 25 different power spectrum shapes by uniformly sampling x_1 and x_2 between -2 and 2. Each power spectrum shape was then scaled to a power equal to 0.5, 1, and 2 times the thermal noise power σ_n^2, resulting in 75 unique mock 21-cm signals. A Gaussian random realisation of a 21-cm signal gridded visibility cube was generated per signal shape and intensity.

(ii) *Injection*: The simulated 21-cm visibility cube was added to the gridded data cube after the outlier flag, just before performing the ML-GPR step.

(iii) *ML-GPR application*: ML-GPR was subsequently applied to the data cube with the injected signal using the same priors as for the original observed data. A new set of optimal foreground, excess and 21-cm kernels was found for each injected signal.

(iv) *Recovered signal estimation*: Residual power spectra were calculated as in Equation (9). A recovered 21-cm power spectrum Δ_{rec}^2 was then estimated by subtracting the residual power spectrum of the original data (without the injected signal) from the residual power spectrum of the data with the injected signal.

(v) *Comparison*: Each recovered 21-cm power spectrum was compared to the power spectrum of the corresponding injected signal Δ_{inj}^2 using the two statistical estimators:

- *z-score*: It describes the deviation of the recovered signal from the injected signal in units of standard deviations at each k-bin. The z-score was estimated as the inverse cumulative distribution function of the normal distribution, given a p-value for each k-bin calculated as the fraction of realisations where $\Delta_{\mathrm{rec}}^2(k) > \Delta_{\mathrm{inj}}^2(k)$ (our null hypothesis). With this null hypothesis, z-score < 0 indicates signal absorption, with z-score < -2 suggesting a suppression beyond the 2σ upper limit.

Figure 15. Box plots of the z-scores (top) and bias factor $b_k = \Delta^2_{\text{rec}}/\Delta^2_{\text{inj}}$ (bottom) as a function of k-mode, resulting from the injection test. Each box represents the inter-quartile range (IQR), with the median indicated by a black solid line, the whiskers extending to 1.5 IQR from the box, and outliers shown as circles. We injected 25 mock 21-cm signals of different shapes, and scaled their intensities to match 0.5 (blue), 1 (orange), and 2 (green) times the noise variance. In the top panel, each box plot shows the z-score distributions for all the 25 shapes, with different colours for the different intensities. The dark and light grey shaded areas indicate the 1σ and 2σ levels, respectively. In the bottom panel, each box plot represents the distribution of the bias factor for all the 75 injected signals.

- Bias: It measures the factor by which the recovered signal differs from the injected signal at each k-bin, calculated as

$$b_k = \frac{\Delta^2_{\text{rec}}(k)}{\Delta^2_{\text{inj}}(k)} . \tag{15}$$

  A z-score < 0 means $b_k < 1$, and the employed ML-GPR model might bring signal loss.

The z-scores for all the 75 signal injection tests are shown in the top panel of Fig. 15. Our ML-GPR model was not able to recover most of the injected signals, resulting in approximately 62, 47, and 24 per cent of z-scores < -2 for $\sigma^2_{21}/\sigma^2_{\text{n}} = 0.5$, 1, and 2, respectively. Most of these are at $0.11 \le k \le 0.27\,h\,\text{cMpc}^{-1}$, while at the edges of the samples k-space the median z-score for all the intensities is within the 2σ upper limit. We also observe a trend with the intensity, where higher σ^2_{21} resulted in better recovered signals. This is shown in Fig. 16 with the power spectra of the different injected signal shapes. The colours indicate the mean z-score over k for each combination of x_1 and x_2. Most of the cases with z-scores < -2 show a flat power spectrum or an upturn at low k. These are more easily absorbed into the mode-mixing and excess components because of their larger spectral coherence scale. Steeper power spectra or with

a bump around $0.2\,h\,\text{cMpc}^{-1}$ were better recovered, being less similar to a foreground component. The fact that injected signals with higher variance were better recovered might also mean that ML-GPR is not successful in modelling data with strong foreground residuals. Because of the limited spatial extent of our sky model and lack of extended emission, our ML-GPR input data have high power close to the primary beam null and in the sidelobes (see Fig. 5, Fig. 8, and Fig. 10). This brought our ML-GPR model to be able to differentiate between the 21-cm signal and the mode-mixing component only when the injected signal is higher than the mode-mixing for most of the k-bins.

The bottom panel of Fig. 15 shows the bias factor b_k for each k-bin. While a few tests returned $b_k \le 1$ at $k = 0.08$ and $0.52\,h\,\text{cMpc}^{-1}$, the median value is below one for all k-bins, as expected from the z-scores. This bias indicates that our ML-GPR model would likely suppress the 21-cm signal, especially when it is at or below the noise level. Therefore, a bias $b_k < 1$ results in lower upper limits, potentially leading to incorrect interpretation. To correct for this bias, we can divide the final power spectrum results by b_k at each k-bin, giving conservative upper limits.

6.3 Data and model power spectra

In addition to the strong foreground residuals, another possible reason for larger biases could be GPR model incompleteness. If there is an extra component not modelled by our set of kernels, adding a signal on top of it might lead to a mix of this component with the injected signal, preventing full recovery of the latter. To ensure this is not the case, we estimated the 'model residual' as

$$\mathbf{m}_{\text{res}} = \mathbb{E}[\mathbf{f}_{\text{ex}}] + \mathbb{E}[\mathbf{f}_{21}] , \tag{16}$$

and compared its power spectrum with the noise-subtracted power spectrum of the data residuals, given by

$$P_{\text{d}} = P_{\text{I,res}} - P_{\text{n}}. \tag{17}$$

If the ML-GPR model is complete and our kernel set fully describes the data, these two estimators should match. The cylindrical power spectra of the data and model residuals are shown in the first and second panels of Fig. 17, respectively. The data power spectrum is obtained by subtracting the thermal noise power spectrum from the ML-GPR residual data. This operation results in some ($k_\perp$-$k_\parallel$)-cells with negative power, particularly in the EoR window, where the residual data variance is dominated by the noise variance (see Fig. 14).

The ratio of the two power spectra is plotted in the third panel of Fig. 17. In the bottom-left corner, where the excess component dominates the residuals, the ratio is around 1. At higher $k_\perp$, within the wedge up to the Cas A direction, the data residuals show more power than the model residuals. Above the Cas A delay lines, there are regions where the model residuals are stronger, but these mainly correspond to the ($k_\perp$-$k_\parallel$)-cells where the noise variance dominates over the excess. The ratio suggests that a component with higher power than the excess at $k_\perp > 0.12\,h\,\text{cMpc}^{-1}$ and within the wedge might be necessary. We tested an excess kernel with $\alpha_{\sigma,\text{ex}} = 0$, and the resulting evidence was lower than for $\alpha_{\sigma,\text{ex}} = -0.25$, indicating that our excess kernel did not better fit the data with higher variance at longer baselines. However, Mertens et al. (2025) showed that using two mode-mixing components with different θ_{mix} improved results at $z = 10.1$, where the foreground is brighter. We do not exclude the possibility that adding an extra excess or mode-mixing component with higher variance at larger $k_\perp$ could reduce the difference between the data and model residuals and ultimately allow the injection tests

Figure 16. Spherical power spectra of the 25 injected 21-cm signal shapes. From left to right, the panels show the same signal shapes with different intensities, namely 0.5, 1, and 2 σ_n^2. Colours indicate the mean z-score, averaged over the k-modes.

Figure 17. Cylindrical power spectra of the noise-subtracted data residuals (first panel), model residuals (second panel), and ratio between them (third panel). The same colour bar is used for the first and second panels. In all panels, the black solid line indicates the horizon delay line, while the black dashed lines indicate the delay ranges where we expect most of the power of Cas A.

to pass with z-score > -2. Nonetheless, since the two estimators agree within their uncertainties (see the spherical power spectra in Section 7), we conclude that our ML-GPR model is mostly complete in the k-bins of interest, with no significant indication of missing components.

7 RESULTS AND UPPER LIMITS

We finally derived the upper limits on the 21-cm signal from the ML-GPR Stokes I residual data. The spherical power spectrum is estimated in seven k-bins logarithmically spaced between 0.06 and 0.5 $h\,\mathrm{cMpc}^{-1}$, with a bin size of $\Delta k/k \approx 0.3$. Similar to Mertens et al. (2020), we subtracted the noise bias from the Stokes I residual power spectrum $\Delta_{\mathrm{I,res}}^2$. We used the time-differenced Stokes V, the same as given as input in the ML-GPR model, to estimate the noise bias power spectrum Δ_n^2. The spherical noise-subtracted power spectrum of the residual is then defined from Equation (17) as

$$\Delta_{21,\mathrm{d}}^2 = \Delta_{\mathrm{I,res}}^2 - \Delta_n^2. \tag{18}$$

Similar to Section 6.3, we compare $\Delta_{21,\mathrm{d}}^2$ with the spherical power spectrum of the model residual $\Delta_{21,\mathrm{m}}^2$, estimated from the visibility cube of Equation (16). The uncertainties of the power spectra are estimated by using a sampling approach within our ML-GPR framework. By generating multiple realizations of the power spectrum through sampling of the hyper-parameter posterior distributions, we obtained a distribution for each k-bin, as described in Section 5.3. We then computed the 97.5$^{\mathrm{th}}$ percentile $\mathcal{P}_{97.5}$ of these distributions, providing the 2σ upper limit at each k-bin.

The resulting power spectra are shown in Fig. 18 and the values reported in Table 3. The comparison between $\Delta_{21,\mathrm{d}}^2$ (orange points) and $\Delta_{21,\mathrm{m}}^2$ (grey points) leads to the same conclusions as in Section 6.3, where no additional GPR component appeared necessary, except at $k = 0.242\,h\,\mathrm{cMpc}^{-1}$, where $\Delta_{21,\mathrm{m}}^2 < \Delta_{21,\mathrm{d}}^2$. No further investigation was required as the measurements are consistent within the 2σ uncertainties. Given the potential signal loss indicated by the injection test, we corrected the noise-subtracted power spectra of the residual data by dividing $\Delta_{21,\mathrm{d}}^2$ by the bias factor b_k at each k-bin. The bias-corrected power spectra (green points) are higher than $\Delta_{21,\mathrm{d}}^2$ at $k < 0.3\,h\,\mathrm{cMpc}^{-1}$, but they become comparable at smaller scales,

Table 3. Spherical power spectrum values at each k-bin for the different 21-cm signal estimators, namely the noise-subtracted residual data $\Delta^2_{21,\mathrm{d}}$, the model residual $\Delta^2_{21,\mathrm{m}}$, and the bias-corrected residual data $\Delta^2_{21,\mathrm{d}}/b_k$, that we consider for our final results. The upper limits are reported as the 97.5^{th} percentile of the power spectrum distributions. The last two columns report the power spectrum of the thermal noise and its 2σ uncertainty.

k $(h\,\mathrm{cMpc}^{-1})$	$\Delta^2_{21,\mathrm{d}}$ (mK^2)	$\mathcal{P}_{97.5}(\Delta^2_{21,\mathrm{d}})$ (mK^2)	$\Delta^2_{21,\mathrm{m}}$ (mK^2)	$\mathcal{P}_{97.5}(\Delta^2_{21,\mathrm{m}})$ (mK^2)	$\mathcal{P}_{97.5}(\Delta^2_{21,\mathrm{d}}/b_k)$ (mK^2)	Δ^2_{n} (mK^2)	$2\Delta^2_{\mathrm{n,err}}$ (mK^2)
0.078	$(89.14)^2$	$(100.99)^2$	$(87.26)^2$	$(98.72)^2$	$(146.61)^2$	$(36.33)^2$	$(22.40)^2$
0.104	$(131.73)^2$	$(145.16)^2$	$(129.05)^2$	$(141.66)^2$	$(208.05)^2$	$(60.48)^2$	$(31.14)^2$
0.134	$(186.67)^2$	$(201.47)^2$	$(177.21)^2$	$(189.17)^2$	$(271.89)^2$	$(97.06)^2$	$(42.15)^2$
0.183	$(263.24)^2$	$(277.86)^2$	$(246.38)^2$	$(260.57)^2$	$(343.33)^2$	$(165.26)^2$	$(59.61)^2$
0.242	$(353.31)^2$	$(372.23)^2$	$(317.65)^2$	$(333.16)^2$	$(428.38)^2$	$(303.60)^2$	$(87.07)^2$
0.320	$(468.72)^2$	$(489.32)^2$	$(450.48)^2$	$(466.60)^2$	$(538.21)^2$	$(442.66)^2$	$(127.87)^2$
0.430	$(640.79)^2$	$(672.55)^2$	$(598.53)^2$	$(618.12)^2$	$(706.98)^2$	$(675.26)^2$	$(185.62)^2$

Figure 18. Final Stokes I spherical power spectra after ML-GPR. We plot both the noise-subtracted power spectrum of the data residuals (orange points) and the power spectrum of the model residuals (grey points). The green circles represent the bias-corrected power spectrum, which is used to estimate the upper limits on the 21-cm signal. The thermal noise level is shown with the blue line. The horizontal error bars indicate the k-bin extension, while the vertical ones and the blue shaded area indicate the 95 per cent confidence interval (i.e. 2σ uncertainties). A small offset in k has been added to the grey points and green circles to avoid overlapping of error bars. The black dashed line represents the lowest achievable upper limit (given the injection test bias) if the ML-GPR residuals were fully thermal noise dominated, namely the 2σ uncertainties of the thermal noise.

where the injection tests mostly returned z-score > -2. For reporting upper limits, we consider the bias-corrected power spectra from the Stokes I data residuals as our final results, following a conservative approach, so that $\Delta^2_{21} = \Delta^2_{21,\mathrm{d}}/b_k$.

The deepest (bias-corrected) upper limit from the 3C 196 field at $z = 9.16$ is $\Delta^2_{21} < (146.61\,\mathrm{mK})^2$ at $k = 0.078\,h\,\mathrm{cMpc}^{-1}$, which is still more than 40 times higher than the achievable upper limit if the residuals were fully consistent with the thermal noise[16]. This is largely due to the bias correction and the inclusion of the excess

---

[16] Subtracting the noise bias from $\Delta_{\mathrm{I,res}}$, if this is consistent with the thermal

variance in the upper limits. At low k, the residual power spectrum is dominated by the excess and subtracting the noise bias, which is an order of magnitude lower, does not significantly reduce the upper limits. Conversely, at higher k, where the excess power is lower, the residuals are noise-dominated. Here, the difference from the deepest achievable upper limits decreases to a factor of 14 at $k = 0.43\,h\,\mathrm{cMpc}^{-1}$, compared to a factor of 100 at the same k in Mertens et al. (2020).

8 DISCUSSION AND CONCLUSIONS

For the first time, we have used LOFAR observations of the 3C 196 field to set upper limits on the 21-cm signal from the EoR. Previous LOFAR 21-cm signal limits were all set using the NCP target field. The 3C 196 field, being closer to the zenith than the NCP, can be observed for only half of the time spent on the NCP but reaches, in principle, the same power spectrum sensitivity. In this work, only a 6-hour night of 3C 196 data has been processed. The main conclusions of our analysis are the following:

New 21-cm signal processing pipeline: The processing is largely similar between the 3C 196 and NCP fields, but here, instead of SAGECAL, we used DP3, which is the standard tool for processing LOFAR data. The most notable difference between the two lies in the spectral smoothing of the DD-gain solutions. While SAGECAL employs a consensus optimisation algorithm to spectrally constrain the solutions using a third-order Bernstein polynomial, the DDECAL tool of DP3 applies a Gaussian smoothing kernel to regularise between iterations. Gan et al. (2023) demonstrated that using a kernel size of 4 MHz yields power spectrum results comparable to those obtained with SAGECAL on the same dataset; therefore, we adopted the same setting.

Sky modelling and source clustering: While the NCP model contains more sources than the 3C 196 sky model, its creation involved significantly more manual work (e.g. Bernardi et al. 2010; Yatawatta et al. 2013) compared to the automatic procedure used here, which is essentially based on a multi-scale and multi-frequency deconvolution done with WSCLEAN on previously DI calibrated data. Moreover, the NCP model was based on apparent flux densities, whereas in this work, we consistently used intrinsic flux densities. This facilitates the use of our model when the field is observed at different local sidereal times. Another notable difference in this work includes the use of

noise, the 2σ upper limits would correspond to the 2σ uncertainties of the noise.

fewer directions in the direction-dependent calibration step (48 for 3C 196 compared to just over 100 for NCP).

Difference in excess power and systematics to NCP field: Having a second field to compare with opens new insights into the cause of the excess power. One of the most striking results from this work is the different behaviour and level of the excess variance in the 3C 196 field compared to the NCP field. The excess frequency coherence and baseline behaviour are different from the excess observed in the NCP observations. Compared to the NCP field, the 3C 196 field has notably different characteristics:

- The presence of a very bright FR-II source (i.e. 3C 196) of 83 Jy at 150 MHz in the centre of the field, compared to the NCP brightest source of ≈ 5 Jy.
- It is observed by LOFAR near zenith, leading to higher sensitivity but larger beam variations due to tracking, while the NCP field is observed at a constant zenith angle of $37°$.
- Tracking the field may reduce the impact of stationary RFI by time-smearing static interference, while this does not occur in NCP observations, where stationary RFI adds up coherently.

A more subtle factor may be the variation of the apparent brightness of sources in the primary beam sidelobes, particularly the four dominant northern A-team sources (i.e. Cas A, Cyg A, Tau A and Vir A). The combination of all these differences makes it hard to attribute the observed differences to a single cause as of yet, but the excess power in the 3C 196 results shows a clear wedge-like structure in the cylindrical power spectrum, and an increase in power towards smaller k-values. This suggests that its origin is a residual foreground modulated by the time-varying primary beam, rather than, for example, interference.

Bias factor from signal injection test: We used GPR for residual foreground removal, which was not used by Patil et al. (2017) but was implemented in more recent NCP analyses (Mertens et al. 2020, 2025). The signal injection test to validate our ML-GPR model returned a median bias factor b_k smaller than one for all the k-bins, with $b_k \approx 0.5$ at $k = 0.078\, h\, \text{cMpc}^{-1}$. Such a bias is not measured in the NCP field and may result from a less complete sky model, which covers just the main lobe of the primary beam. In this case, such a bias could be removed by constructing a deeper and more spatially extended sky model.

New upper limits: We set a best (bias-corrected) upper limit of $\Delta_{21}^2 < (146.61\, \text{mK})^2$ at $k = 0.078\, h\, \text{cMpc}^{-1}$ and $z = 9.16$ from the 3C 196 field. This is approximately seven times higher than the current deepest upper limit from the NCP field at the same k-bin and redshift (Mertens et al. 2025). However, this difference is expected, given that our result is based on a single 6-hour night, whereas the NCP upper limits are derived from 141 hours of data. A more appropriate comparison is with Patil et al. (2017), which analysed a single 13-hour night of NCP observations. The longer integration time compensates for the higher noise variance at the NCP, due to its greater distance from the zenith, and thus we expect similar upper limits. Although the k-bins do not match exactly, Patil et al. (2017) report an upper limit of $\Delta_{21}^2 < (296.1\, \text{mK})^2$ at $k = 0.1\, h\, \text{cMpc}^{-1}$, whereas our limit at the same k is more than a factor of two lower, with $\Delta_{21}^2 < (208.05\, \text{mK})^2$. This highlights the potential of the 3C 196 field to provide deeper upper limits than the NCP field, despite our conservative approach of correcting for the GPR bias. Even stronger constraints could be achieved in the future if this bias is reduced or eliminated.

In conclusion, our results suggest that the 3C 196 field has lower systematics and could achieve deeper upper limits than the NCP field when a larger number of nights is processed, potentially providing competitive constraints on the 21-cm signal. One of the advantages of having two independent deep fields is that their power spectra can be combined incoherently, reducing the upper limits by a factor of $\sqrt{2}$. Further comparisons between the two fields will help identify the origin of the excess power and refine observational strategies for future experiments.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

Abdurashidova Z., et al., 2022, ApJ, 925, 221
Acharya A., et al., 2023, MNRAS,
Acharya A., et al., 2024, MNRAS, 527, 7835
Arrabal Haro P., et al., 2023, Nature, 622, 707
Arras P., Reinecke M., Westermann R., Enßin T. A., 2021, A&A, 646, A58
Asad K. M. B., et al., 2015, MNRAS, 451, 3709
Astropy Collaboration et al., 2022, ApJ, 935, 167
Barkana R., Loeb A., 2001, Phys. Rep., 349, 125
Barry N., Hazelton B., Sullivan I., Morales M. F., Pober J. C., 2016, MNRAS, 461, 3135
Beardsley A. P., et al., 2016, ApJ, 833, 102
Becker R. H., et al., 2001, AJ, 122, 2850
Bernardi G., et al., 2009, A&A, 500, 965
Bernardi G., et al., 2010, A&A, 522, A67
Bianco M., Giri S. K., Prelogović D., Chen T., Mertens F. G., Tolley E., Mesinger A., Kneib J.-P., 2024, MNRAS, 528, 5212

Bobin J., Starck J.-L., Fadili J., Moudden Y., 2007, IEEE Transactions on Image Processing, 16, 2662

Bobin J., Moudden Y., Starck J. L., Fadili J., Aghanim N., 2008, Statistical Methodology, 5, 307

Bowman J. D., Morales M. F., Hewitt J. N., 2009, ApJ, 695, 183

Boylan-Kolchin M., 2023, Nature Astronomy, 7, 731

Brackenhoff S. A., et al., 2024, MNRAS, 533, 632

Brackenhoff S. A., et al., 2025, arXiv e-prints, p. arXiv:2504.02483

Bradley L. D., et al., 2023, ApJ, 955, 13

Brentjens M. A., Mol J. D., 2018, in Astrophysics and Space Science Library. p. 19, doi:10.1007/978-3-319-23434-2_2

Buchner J., 2016, Statistics and Computing, 26, 383

Buchner J., 2019, PASP, 131, 108005

Buchner J., 2021, The Journal of Open Source Software, 6, 3001

Carniani S., et al., 2024, Nature, 633, 318

Ceccotti E., et al., 2025, A&A, 696, A56

Chapman E., et al., 2013, MNRAS, 429, 165

Chege J. K., et al., 2024, A&A, 692, A211

Chemerynska I., et al., 2024, MNRAS, 531, 2615

Chokshi A., Barry N., Line J. L. B., Jordan C. H., Pindor B., Webster R. L., 2024, MNRAS, 534, 2475

Cunnington S., Irfan M. O., Carucci I. P., Pourtsidou A., Bobin J., 2021, MNRAS, 504, 208

Datta A., Bowman J. D., Carilli C. L., 2010, ApJ, 724, 526

DeBoer D. R., et al., 2017, PASP, 129, 045001

Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L. W., 2009, IEEE Proceedings, 97, 1482

Eilers A.-C., Davies F. B., Hennawi J. F., 2018, ApJ, 864, 53

Ewall-Wice A., Dillon J. S., Liu A., Hewitt J., 2017, MNRAS, 470, 1849

Fan X., et al., 2006, AJ, 132, 117

Ferrara A., Pallottini A., Dayal P., 2023, MNRAS, 522, 3986

Finkelstein S. L., et al., 2023, ApJ, 946, L13

Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181

Gan H., et al., 2022, A&A, 663, A9

Gan H., et al., 2023, A&A, 669, A20

Gehlot B. K., et al., 2018, MNRAS, 478, 1484

Gehlot B. K., et al., 2021, MNRAS, 506, 4578

Ghara R., Choudhury T. R., Datta K. K., 2015a, MNRAS, 447, 1806

Ghara R., Datta K. K., Choudhury T. R., 2015b, MNRAS, 453, 3143

Ghara R., Mellema G., Giri S. K., Choudhury T. R., Datta K. K., Majumdar S., 2018, MNRAS, 476, 1741

Ghara R., et al., 2020, MNRAS, 493, 4728

Giri S. K., Mellema G., Dixon K. L., Iliev I. T., 2018a, MNRAS, 473, 2949

Giri S. K., Mellema G., Ghara R., 2018b, MNRAS, 479, 5596

Gooch R., 1996, in Jacoby G. H., Barnes J., eds, Astronomical Society of the Pacific Conference Series Vol. 101, Astronomical Data Analysis Software and Systems V. p. 80

Gorce A., et al., 2023, MNRAS, 520, 375

Greig B., et al., 2021, MNRAS, 501, 1

Gupta Y., et al., 2017, Current Science, 113, 707

HERA Collaboration et al., 2023, ApJ, 945, 124

Harris C. R., et al., 2020, Nature, 585, 357

Hothi I., et al., 2021, MNRAS, 500, 2264

Hunter J. D., 2007, Computing in Science and Engineering, 9, 90

Jelić V., et al., 2008, MNRAS, 389, 1319

Jelić V., Zaroubi S., Labropoulos P., Bernardi G., de Bruyn A. G., Koopmans L. V. E., 2010, MNRAS, 409, 1647

Jelić V., et al., 2015, A&A, 583, A137

Joye W. A., Mandel E., 2003, in Payne H. E., Jedrzejewski R. I., Hook R. N., eds, Astronomical Society of the Pacific Conference Series Vol. 295, Astronomical Data Analysis Software and Systems XII. p. 489

Keller P. M., Thyagarajan N., Kumar A., Kanekar N., Bernardi G., 2024, MNRAS, 528, 5692

Kern N. S., Liu A., 2021, MNRAS, 501, 1463

Kern N. S., et al., 2020, ApJ, 888, 70

Kolopanis M., Pober J. C., Jacobs D. C., McGraw S., 2023, MNRAS, 521, 5120

Konno A., et al., 2014, ApJ, 797, 16

Koopmans L. V. E., 2010, ApJ, 718, 963

Koopmans L., et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 1 (arXiv:1505.07568), doi:10.22323/1.215.0001

Liu A., Shaw J. R., 2020, PASP, 132, 062001

Liu A., Parsons A. R., Trott C. M., 2014a, Phys. Rev. D, 90, 023018

Liu A., Parsons A. R., Trott C. M., 2014b, Phys. Rev. D, 90, 023019

Loeb A., Furlanetto S. R., 2013, The First Galaxies in the Universe

Lonsdale C. J., et al., 2009, IEEE Proceedings, 97, 1497

Mason C. A., Trenti M., Treu T., 2023, MNRAS, 521, 497

Mazumder A., Datta A., Chakraborty A., Majumdar S., 2022, MNRAS, 515, 4020

McKinney W., 2010, in van der Walt S., Millman J., eds, Proceedings of the 9th Python in Science Conference. pp 56–61, doi:10.25080/Majora-92bf1922-00a

McLeod D. J., et al., 2024, MNRAS, 527, 5004

Mertens F. G., Ghosh A., Koopmans L. V. E., 2018, MNRAS, 478, 3640

Mertens F. G., et al., 2020, MNRAS, 493, 1662

Mertens F. G., Bobin J., Carucci I. P., 2024, MNRAS, 527, 3517

Mertens F. G., et al., 2025, arXiv e-prints, p. arXiv:2503.05576

Mevius M., et al., 2016, Radio Science, 51, 927

Mevius M., et al., 2022, MNRAS, 509, 3693

Mitchell D. A., Greenhill L. J., Wayth R. B., Sault R. J., Lonsdale C. J., Cappallo R. J., Morales M. F., Ord S. M., 2008, IEEE Journal of Selected Topics in Signal Processing, 2, 707

Mondal R., et al., 2020, MNRAS, 498, 4178

Morales M. F., Hewitt J., 2004, ApJ, 615, 7

Morales M. F., Hazelton B., Sullivan I., Beardsley A., 2012, ApJ, 752, 137

Morales M. F., Beardsley A., Pober J., Barry N., Hazelton B., Jacobs D., Sullivan I., 2019, MNRAS, 483, 2207

Mouri Sardarabadi A., Koopmans L. V. E., 2019, MNRAS, 483, 5480

Munshi S., et al., 2024, A&A, 681, A62

Munshi S., et al., 2025a, arXiv e-prints, p. arXiv:2503.21728

Munshi S., et al., 2025b, A&A, 693, A276

Offringa A. R., 2016, A&A, 595, A99

Offringa A. R., Smirnov O. M., 2017, MNRAS, 471, 301

Offringa A. R., de Bruyn A. G., Biehl M., Zaroubi S., Bernardi G., Pandey V. N., 2010, MNRAS, 405, 155

Offringa A. R., van de Gronde J. J., Roerdink J. B. T. M., 2012, A&A, 539, A95

Offringa A. R., et al., 2013, A&A, 549, A11

Offringa A. R., et al., 2014, MNRAS, 444, 606

Offringa A. R., et al., 2016, MNRAS, 458, 1057

Offringa A. R., Mertens F., Koopmans L. V. E., 2019a, MNRAS, 484, 2866

Offringa A. R., Mertens F., van der Tol S., Veenboer B., Gehlot B. K., Koopmans L. V. E., Mevius M., 2019b, A&A, 631, A12

Ouchi M., et al., 2010, ApJ, 723, 869

Paciga G., et al., 2013, MNRAS, 433, 639

Pandey V. N., Koopmans L. V. E., Tiesinga E., Albers W., Koers H. U. A., 2020, in Pizzo R., Deul E. R., Mol J. D., de Plaa J., Verkouter H., eds, ASP Conf. Ser. Vol. 527, ADASS XXIX.

Parsons A. R., Backer D. C., 2009, AJ, 138, 219

Parsons A. R., Pober J. C., Aguirre J. E., Carilli C. L., Jacobs D. C., Moore D. F., 2012, ApJ, 756, 165

Patil A. H., et al., 2016, MNRAS, 463, 4317

Patil A. H., et al., 2017, ApJ, 838, 65

Planck Collaboration et al., 2016, A&A, 594, A13

Planck Collaboration et al., 2020, A&A, 641, A6

Pritchard J. R., Loeb A., 2012, Reports on Progress in Physics, 75, 086901

Rahimi M., et al., 2021, MNRAS, 508, 5954

Rath E., et al., 2024, arXiv e-prints, p. arXiv:2406.08549

Salvini S., Wijnholds S. J., 2014, A&A, 571, A97

Santos M. G., Cooray A., Knox L., 2005, ApJ, 625, 575

Scaife A. M. M., Heald G. H., 2012, MNRAS, 423, L30

Smirnov O. M., Tasse C., 2015, MNRAS, 449, 2668

Spinelli M., Bernardi G., Santos M. G., 2018, MNRAS, 479, 275

Stein M. L., 1999, Interpolation of Spatial Data: Some Theory for Kriging. Springer Series in Statistics, Springer-Verlag,

New York, doi:10.1007/978-1-4612-1494-6, http://dx.doi.org/10.1007/978-1-4612-1494-6

Taylor A. J., Cowie L. L., Barger A. J., Hu E. M., Songaila A., 2021, ApJ, 914, 79

Thyagarajan N., et al., 2015, ApJ, 804, 14

Trott C. M., et al., 2020, MNRAS, 493, 4711

Vedantham H. K., Koopmans L. V. E., 2015, MNRAS, 453, 925

Vedantham H. K., Koopmans L. V. E., 2016, MNRAS, 458, 3099

Vedantham H., Udaya Shankar N., Subrahmanyan R., 2012, ApJ, 745, 176

Venemans B. P., et al., 2013, ApJ, 779, 24

Virtanen P., et al., 2020, Nature Methods, 17, 261

Wilensky M. J., Morales M. F., Hazelton B. J., Barry N., Byrne R., Roy S., 2019, PASP, 131, 114507

Witstok J., et al., 2024, A&A, 682, A40

Yatawatta S., 2015, MNRAS, 449, 4506

Yatawatta S., 2016, arXiv e-prints, p. arXiv:1605.09219

Yatawatta S., 2018, arXiv e-prints, p. arXiv:1805.00265

Yatawatta S., et al., 2013, A&A, 550, A136

Yatawatta S., Diblen F., Spreeuw H., 2017, in 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP). pp 1–5, doi:10.1109/CAMSAP.2017.8313135

Zarka P., Girard J. N., Tagger M., Denis L., 2012, in Boissier S., de Laverny P., Nardetto N., Samadi R., Valls-Gabaud D., Wozniak H., eds, SF2A-2012: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics. pp 687–694

Zheng Z.-Y., et al., 2017, ApJ, 842, L22

de Gasperin F., et al., 2019, A&A, 622, A5

van Diepen G., Dijkema T. J., Offringa A., 2018, DPPP: Default Pre-Processing Pipeline, Astrophysics Source Code Library, record ascl:1804.003 (ascl:1804.003)

van Haarlem M. P., et al., 2013, A&A, 556, A2

Table A1. Details of the 3C 196 field 'DD sky model' used in the EoR processing pipeline. For each cluster, we provide the number of clean components, central coordinates, distance from the field centre, total flux density of the components at 150 MHz, and the most notable dominating sources. Cluster 48 contains only Cas A.

Cluster	Components	Right Ascension (J2000)	Declination (J2000)	Distance (deg)	Flux (Jy)	Bright sources
1	1864	$08^h13^m35\overset{s}{.}40$	$+48°12'52\overset{''}{.}15$	0.00	84.84	3C 196
2	205	$07^h59^m28\overset{s}{.}42$	$+50°37'59\overset{''}{.}52$	3.33	17.68	J0801
3	105	$08^h20^m11\overset{s}{.}05$	$+46°48'59\overset{''}{.}74$	1.79	14.76	3C 197.1
4	142	$08^h14^m46\overset{s}{.}46$	$+46°32'10\overset{''}{.}76$	1.69	10.76	4C+46.17
5	148	$08^h04^m45\overset{s}{.}87$	$+48°12'42\overset{''}{.}03$	1.47	10.24	4C+48.21, J0805
6	230	$08^h04^m41\overset{s}{.}64$	$+47°14'36\overset{''}{.}87$	1.78	10.17	4C+47.27
7	232	$08^h32^m01\overset{s}{.}61$	$+46°38'33\overset{''}{.}93$	3.49	9.67	4C+45.16
8	251	$08^h21^m34\overset{s}{.}60$	$+51°18'22\overset{''}{.}39$	3.35	9.21	
9	319	$08^h00^m13\overset{s}{.}09$	$+47°43'30\overset{''}{.}54$	2.29	8.83	
10	297	$08^h16^m45\overset{s}{.}97$	$+44°52'18\overset{''}{.}25$	3.39	8.03	
11	272	$08^h02^m28\overset{s}{.}17$	$+46°16'22\overset{''}{.}23$	2.71	7.85	4C+45.14
12	307	$08^h24^m39\overset{s}{.}05$	$+45°24'46\overset{''}{.}95$	3.38	7.61	
13	180	$08^h05^m46\overset{s}{.}53$	$+49°45'57\overset{''}{.}56$	2.01	7.41	
14	300	$08^h09^m12\overset{s}{.}23$	$+50°32'58\overset{''}{.}73$	2.44	7.07	
15	261	$08^h15^m23\overset{s}{.}47$	$+50°22'49\overset{''}{.}12$	2.19	6.99	
16	287	$08^h10^m15\overset{s}{.}24$	$+45°22'20\overset{''}{.}55$	2.90	6.62	
17	168	$08^h10^m11\overset{s}{.}40$	$+49°24'27\overset{''}{.}33$	1.32	6.71	4C+49.17
18	334	$08^h14^m40\overset{s}{.}70$	$+51°26'14\overset{''}{.}58$	3.23	6.42	
19	260	$08^h32^m51\overset{s}{.}59$	$+49°10'00\overset{''}{.}74$	3.32	5.86	
20	282	$08^h19^m44\overset{s}{.}00$	$+50°00'01\overset{''}{.}06$	2.05	5.79	
21	206	$08^h19^m37\overset{s}{.}25$	$+45°58'18\overset{''}{.}25$	2.47	5.22	
22	289	$08^h30^m41\overset{s}{.}58$	$+50°18'17\overset{''}{.}88$	3.49	4.95	
23	255	$08^h25^m54\overset{s}{.}36$	$+46°32'59\overset{''}{.}68$	2.67	4.74	
24	198	$08^h27^m31\overset{s}{.}48$	$+48°33'38\overset{''}{.}99$	2.34	4.65	
25	183	$08^h25^m53\overset{s}{.}42$	$+47°35'37\overset{''}{.}47$	2.15	4.75	
26	237	$08^h16^m09\overset{s}{.}52$	$+49°21'40\overset{''}{.}51$	1.22	4.67	
27	192	$08^h02^m32\overset{s}{.}20$	$+48°49'26\overset{''}{.}63$	1.93	4.55	
28	296	$08^h07^m45\overset{s}{.}51$	$+51°26'56\overset{''}{.}72$	3.37	4.44	
29	377	$08^h04^m24\overset{s}{.}96$	$+45°10'35\overset{''}{.}09$	3.42	4.11	
30	207	$08^h21^m35\overset{s}{.}04$	$+49°11'41\overset{''}{.}45$	1.64	4.17	
31	125	$08^h18^m55\overset{s}{.}24$	$+47°41'38\overset{''}{.}58$	1.03	3.98	
32	308	$07^h53^m23\overset{s}{.}72$	$+47°32'37\overset{''}{.}77$	3.45	3.96	
33	142	$08^h14^m05\overset{s}{.}45$	$+47°33'27\overset{''}{.}66$	0.66	3.89	
34	288	$07^h53^m30\overset{s}{.}27$	$+49°13'14\overset{''}{.}77$	3.46	3.86	
35	313	$08^h33^m49\overset{s}{.}98$	$+48°03'04\overset{''}{.}84$	3.38	3.77	
36	302	$07^h56^m33\overset{s}{.}12$	$+48°30'42\overset{''}{.}17$	2.85	3.84	
37	120	$08^h10^m42\overset{s}{.}13$	$+48°38'39\overset{''}{.}80$	0.64	3.88	
38	246	$08^h09^m27\overset{s}{.}78$	$+46°32'00\overset{''}{.}40$	1.82	3.23	
39	213	$08^h00^m02\overset{s}{.}18$	$+49°34'00\overset{''}{.}91$	2.61	3.21	
40	186	$08^h22^m29\overset{s}{.}39$	$+48°30'05\overset{''}{.}16$	1.51	3.02	
41	124	$08^h08^m50\overset{s}{.}00$	$+48°09'15\overset{''}{.}96$	0.80	2.70	
42	302	$07^h57^m35\overset{s}{.}33$	$+46°17'19\overset{''}{.}42$	3.33	2.67	
43	248	$08^h24^m32\overset{s}{.}11$	$+50°29'43\overset{''}{.}99$	2.89	2.09	
44	110	$08^h15^m43\overset{s}{.}33$	$+48°34'08\overset{''}{.}63$	0.50	2.03	
45	184	$08^h09^m29\overset{s}{.}40$	$+47°28'17\overset{''}{.}88$	1.01	1.95	
46	202	$08^h26^m12\overset{s}{.}20$	$+49°29'50\overset{''}{.}70$	2.44	1.88	
47	110	$08^h19^m03\overset{s}{.}97$	$+48°20'07\overset{''}{.}44$	0.93	1.35	
48	10	$23^h23^m24\overset{s}{.}23$	$+58°48'49\overset{''}{.}79$	66.14	11 734.28	Cas A

APPENDIX A: DIRECTION-DEPENDENT SKY MODEL

In Table A1, we provide information on each of the 48 clusters that constitute the 3C 196 'DD sky model' created in Section 3.4.

Figure B1. Calibration solutions obtained from the DI spectral smooth calibration of the 3C 196 field. The amplitude of the station gain g as a function of frequency is shown on the left, and the power spectrum of gg^* as a function of delay is shown on the right. Each row corresponds to one element of the 2×2 Jones matrix. Time-averaged solutions are shown for each core station (light blue) and remote station (dark blue), with the mean over all stations shown in red.

APPENDIX B: CALIBRATION SOLUTIONS

Here, we provide a summary of the calibration solutions obtained from running the EoR pipeline on our 3C 196 field dataset (see Section 4). In each figure, we plot the amplitude of the station gains g as a function of frequency (left) and the power spectrum of gg^* as a function of time delay (right). Figures B1 and B2 show the time-averaged solutions from the DI spectral smooth and bandpass calibrations, respectively. The DD calibration solutions, averaged over time and stations, are shown in Fig. B3 for each of the main field clusters, with different colours indicating the cluster distance from the centre (see Table A1). The DD solutions for the Cas A direction (cluster 48) are presented separately in Fig. B4.
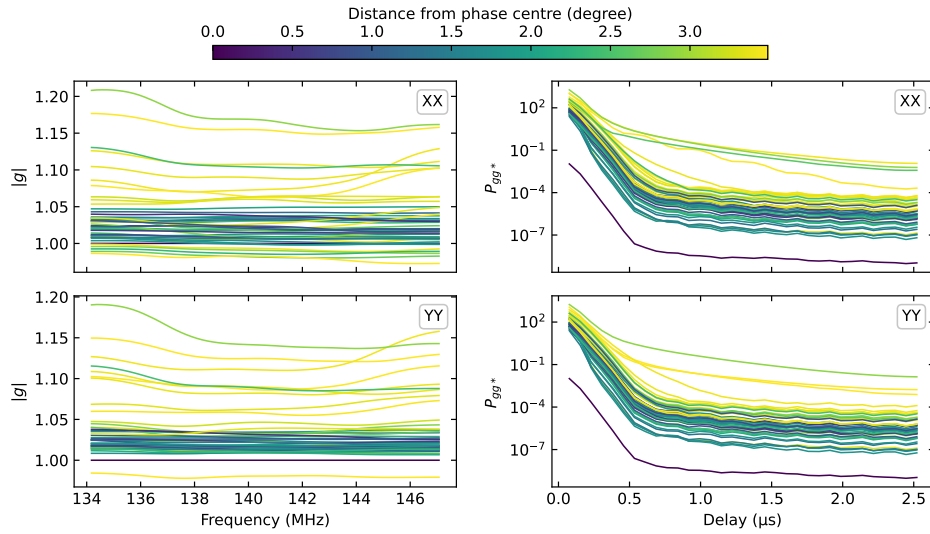
APPENDIX C: LINEARLY POLARIZED INTENSITY MAP

The left and middle panels of Fig. C1 show the Stokes Q and U images at 140.4 MHz of the 3C 196 field after the DD subtraction (Section 4.3). To compare our polarized emission results with Jelić et al. (2015), we calculate the total polarized intensity (PI) at a Faraday depth of $0\,\mathrm{rad/m^2}$ as

$$\mathrm{PI} = \frac{1}{N_{\mathrm{SB}}} \sum_{i=1}^{N_{\mathrm{SB}}} \sqrt{\mathrm{Q}_i^2 + \mathrm{U}_i^2}\,, \tag{C1}$$
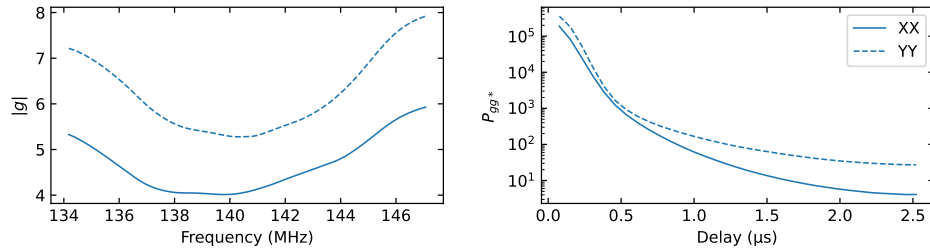
where N_{SB} is the number of individually imaged sub-bands, which in our case is 67 for the redshift bin of interest (corresponding to the 134.2–147.1 MHz frequency range). Here, Q_i and U_i represent the images of each individually imaged sub-band. The obtained PI map only slightly resembles the structures found by Jelić et al. (2015) in Fig. 3. However, a direct comparison is not straightforward due to two key differences: (i) we did not perform any rotation measure (RM) calibration or synthesis, whereas Jelić et al. (2015) corrected for RM using ionospheric data; and (ii) their analysis used a different uv-cut ($10-800\lambda$) compared to ours ($50-500\lambda$), leading to differences in the synthesised beam, sensitivity, and scale of emission captured. Nevertheless, we can still conclude that the 3C 196 field is strongly polarized, and Stokes Q and U may leak power into Stokes V, as observed in Fig. 8.
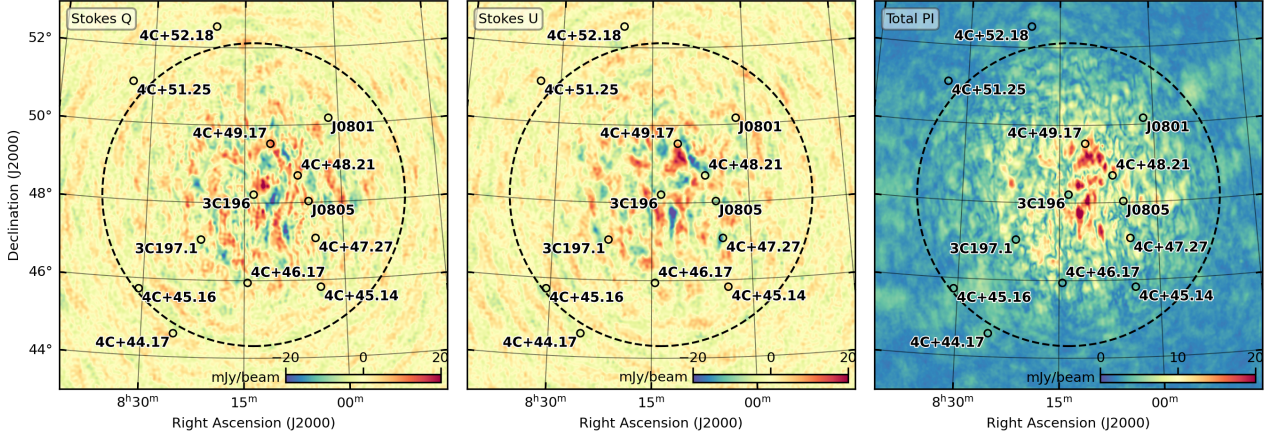
Figure B2. Calibration solutions obtained from the DI bandpass calibration of the 3C 196 field. The amplitude of the station gain g as a function of frequency is shown on the left, and the power spectrum of gg^* as a function of delay is shown on the right. The XX and YY elements of the Jones matrix are shown in the top and bottom rows, respectively. Time-averaged solutions are shown for each core station (light blue) and remote station (dark blue), with the mean over all stations shown in red.

Figure B3. Calibration solutions obtained from the DD calibration of the 3C 196 field. The amplitude of the gain g as a function of frequency is shown on the left, and the power spectrum of gg^* as a function of delay is shown on the right. The XX and YY elements of the Jones matrix are shown in the top and bottom rows, respectively. Each line represents the time and station-averaged gains for each of the 47 main field clusters. The cluster distance from the field centre is colour-coded, ranging from dark blue for 3C 196 to yellow for the farthest cluster (3.49°).

Figure B4. Calibration solutions obtained from the DD calibration of the 3C 196 field for the Cas A direction. The amplitude of the gain g as a function of frequency is shown on the left, and the power spectrum of gg^* as a function of delay is shown on the right. The XX and YY elements of the Jones matrix are shown with solid and dashed lines, respectively. The solutions are averaged over time and stations.

Figure C1. Stokes Q (left) and U (middle) images at 140.4 MHz of the 3C 196 field. The total PI map at a Faraday depth of $0\,\mathrm{rad/m^2}$ is shown in the right panel. All the images have a pixel scale of 0.5 arcmin and a field of view of $10° × 10°$. The uv-plane was gridded with a natural weighting scheme and only baselines between 50 and 500λ were used. The black dashed circle indicates the $3.9°$ radius of the sky model. The brightest sources are highlighted with black and red circles.

Table D1. Summary of kernel shape tests for the excess component in the ML-GPR model for the 3C 196 field. Each test explores different covariance functions, parametrizations, and hyper-parameter priors, as described in Section 5.2. The parameter $\log \mathcal{Z}$ represents the marginal likelihood (evidence), with higher values indicating better model fit.

Parameter	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
Covariance	RBF	RBF	RBF	$\mu = 3/2$	$\mu = 5/2$	$\mu = 5/2$	$\mu = 5/2$
Parametrization	α-coefficient	Wedge	α-coefficient	α-coefficient	α-coefficient	α-coefficient	α-coefficient
Prior $\sigma_{\mathrm{ex}}^2/\sigma_{\mathrm{d}}^2$	$\log \mathcal{U}(-4, -2)$	$\log \mathcal{U}(-4, -2)$	$\log \mathcal{U}(-4, -2)$	$\log \mathcal{U}(-4, -2)$	$\log \mathcal{U}(-4, -2)$	$\log \mathcal{U}(-4, -2)$	$\log \mathcal{U}(-4, -2)$
Prior l_{ex} (MHz)	$\mathcal{U}(0.2, 0.8)$	–	–	–	–	–	–
Prior $\eta_{\mathrm{buffer,ex}}$ (µs)	–	Fixed(2.6)	–	–	–	–	–
Prior θ_{ex} (rad)	–	$\mathcal{U}(0.1, 2)$	–	–	–	–	–
Prior $l_{0,\mathrm{ex}}$ (MHz)	–	–	$\mathcal{U}(0.2, 0.8)$	$\mathcal{U}(0.2, 0.8)$	$\mathcal{U}(0.2, 0.8)$	$\mathcal{U}(0.2, 0.8)$	$\mathcal{U}(0.2, 0.8)$
Prior α_{ex}	–	–	$\mathcal{U}(-5, 60)$	$\mathcal{U}(-5, 60)$	$\mathcal{U}(-5, 60)$	$\mathcal{U}(-5, 60)$	$\mathcal{U}(-5, 60)$
Prior $\sigma_{\alpha,\mathrm{ex}}^2$	–	–	$\mathcal{U}(-1, 1)$	$\mathcal{U}(-1, 1)$	$\mathcal{U}(-1, 1)$	Fixed(0)	Fixed(-0.25)
$\log \mathcal{Z}$	$37\,160.6 \pm 0.7$	$37\,229.5 \pm 0.6$	$37\,320.5 \pm 0.7$	$37\,301.3 \pm 0.8$	$37\,310.7 \pm 0.4$	$37\,305.2 \pm 0.9$	$37\,312.8 \pm 0.6$

APPENDIX D: TESTS FOR EXCESS POWER KERNELS

As described in Section 5.2, most of the effort in constructing the GPR model focused on finding the optimal shape for the excess covariance function. Table D1 summarises the results from tests with different kernel shapes and hyper-parameter priors. Test 7 gave the highest marginal likelihood and was selected for modelling the residual data. Although Test 3 provided an evidence $\log \mathcal{Z}$ higher than Test 7, all the RBF covariances resulted in converging 21-cm signal parameters. This behaviour, unexpected for a component well below the thermal noise level, suggested potential errors in the model fitting. Therefore, we discarded all the RBF tests, but kept them in the table for completeness.

This paper has been typeset from a TEX/LATEX file prepared by the author.